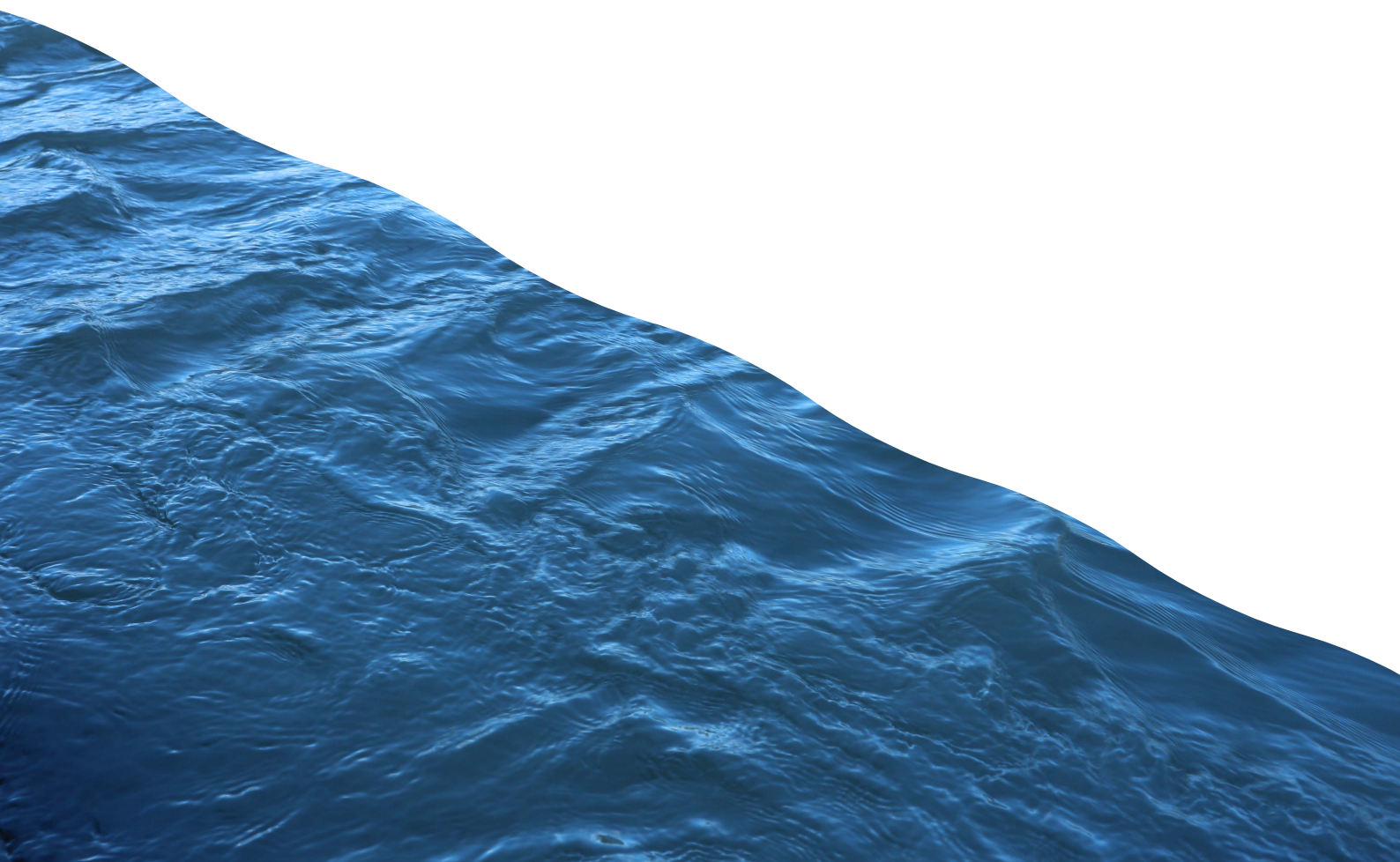
Quality of Crowdsourced Water Level Observations

Barbara Strobl



Quality of Crowdsourced Water Level Observations

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Barbara Strobl

aus

Österreich

Promotionskommission

Prof. Dr. Jan Seibert (Vorsitz)

Dr. Ilja van Meerveld

Prof. Dr. Ross Purves

Zürich, 2020

Für meine Oma

ABSTRACT

Hydrological data are crucial for a better understanding of hydrological processes and can help improve models to predict floods and droughts, to allocate water resources and to better manage hydropower. However, hydrological data are often scarce, as gauging stations are expensive to build and maintain. Citizen science can help fill such data gaps and thereby complement the existing hydrological measurement network. To maximize its use, citizen science in hydrology requires innovative and novel approaches that are aligned with the capabilities and the equipment of citizen scientists. One such approach is a virtual staff gauge, i.e., a sticker with a staff gauge that is inserted onto a photograph of a river that can be used with a mobile smartphone application. The virtual staff gauge is placed on the photograph of the first observation of a particular location, which results in the reference picture. For further observations, citizen scientists compare the current water level to the virtual staff gauge in this reference picture and submit an estimate of the new water level and a new photograph. The water level estimates are measured in classes defined by the staff gauge and have no absolute units.

Compared to streamflow, water level classes are easier for citizen scientists to estimate, and the data are therefore more reliable. However, even for water level class estimates, there is still potential for mistakes, either when placing the virtual staff gauge on the reference picture or when making a new estimate of the water level. Two strategies to reduce these data errors were explored in this thesis. First, data quality control was crowdsourced through a gamified web interface. Citizen scientists can vote on the water level classes that they believe are accurate, based on the comparison of the reference picture and the uploaded photograph. Several votes were collected per observation submitted via the app and the mean value of all votes was used to either confirm or correct the initial water level class estimate that was submitted via the app. This interface provides a scalable way for citizen scientists to check each other's submissions and is therefore also applicable to other large scale citizen science projects.

Second, the gamified interface can also be used for training new citizen scientists. By playing the game, citizen scientists become familiar with the concept of the virtual staff gauge. Voting on a water level class makes them realise which virtual staff gauge placements facilitate or complicate further estimates. This training helps citizen scientists to better place virtual staff gauges in the smartphone application and therefore helps to improve the quality of the reference pictures and all further observations.

This thesis shows that it is possible to crowdsource water level class data through a mobile smartphone application on a global scale. Crowdsourcing data quality control not only results in higher data quality, but also trains new citizen scientists. The crowdsourced water level class data can be used to constrain hydrological models, which can simulate streamflow time series.

ZUSAMMENFASSUNG

Hydrologische Daten sind für ein besseres Verständnis hydrologischer Prozesse essentiell und helfen Hochwasser oder Dürreperioden vorherzusagen, Wasserressourcen zu verteilen und Wasserkraft besser zu nutzen. An vielen Standorten sind jedoch keine hydrologischen Daten vorhanden, da Messstationen teuer zu errichten und zu erhalten sind. Citizen Science (auch Bürgerwissenschaften genannt) kann helfen, diese Datenlücken zu füllen und das hydrologische Messnetz zu ergänzen. Um das Potential von Citizen Science bestmöglich auszunutzen, müssen die Fähigkeiten und Messmöglichkeiten der Citizen Scientists berücksichtigt und innovative Ansätze entwickelt werden; wie zum Beispiel die virtuelle Messlatte. Diese ist eine Art Aufkleber, der digital auf ein Flussfoto geklebt wird und mithilfe einer App weltweit verwendet werden kann. Die virtuelle Messlatte wird an einem Standort eingerichtet und in einem Referenzfoto abgespeichert. Für weitere Beobachtungen am selben Standort bezieht sich der Citizen Scientist immer auf dieses Referenzfoto und fügt dabei ein weiteres Foto sowie eine Schätzung des Wasserstandes hinzu. Der Wasserstand wird in Klassen geschätzt, welche durch die virtuelle Messlatte definiert werden.

Für Citizen Scientists ist es einfacher Wasserstandsklassen als Abfluss abzuschätzen, wodurch die Verlässlichkeit der Daten verbessert wird. Gelegentlich passieren dennoch Fehler; entweder beim Platzieren der virtuellen Messlatte auf dem Referenzfoto oder beim Abschätzen einer weiteren Wasserstandsbeobachtung. Um diese Fehler zu verringern, wurden zwei Strategien entwickelt. Einerseits wird die Datenqualität von vielen Citizen Scientists in einem Spiel kontrolliert, wobei hochgeladene Fotos mehrmals neu klassifiziert werden. Diese Schätzungen werden gemittelt, um den ursprünglich angegebenen Wasserstandswert entweder zu bestätigen oder zu korrigieren.

Andererseits kann das Spiel zusätzlich zur Qualitätskontrolle auch als Training für neue Citizen Scientists verwendet werden. Während des Spielens werden die Citizen Scientists mit der virtuellen Messlatte vertraut und lernen gute und schlechte Platzierungen zu erkennen. Somit hilft das Training neuen Citizen Scientists virtuelle Messlatten in der App besser zu platzieren, was zu verbesserten Referenzfotos und Daten führt.

Diese Arbeit zeigt, dass Schätzungen von Wasserstandsklassen mithilfe einer Smartphone Applikation weltweit gesammelt werden können. Die spielerische Datenqualitätskontrolle hilft zusätzlich neue Citizen Scientists zu trainieren.

PLAIN LANGUAGE SUMMARY

Information about water levels and flow in rivers is important to predict floods and droughts, to help hydropower operations, and to manage water resources. Such information does not always exist, because measurement stations are expensive to build and maintain. New methods, such as citizen science, can help to fill this gap. Data collection methods for citizen science should be as user friendly as possible. One such method is the virtual staff gauge that can be used in an app. With this method, the citizen scientist can add a *"sticker"* of a ruler with classes to a photograph of a river. This photograph is used as a reference by citizen scientists who return to the same place to make a new measurement of the water level.

The work described in this thesis shows that water level classes are relatively easy for citizen scientists to estimate and the collected data is quite reliable. However, sometimes citizen scientists still make mistakes, either when placing the virtual staff gauge on the photograph of a river or when making a new water level class estimate. Therefore, it is important to check how accurate these measurements are. Two different methods to improve the data are presented in this thesis. First, everybody can help to check the data, meaning that the data quality control can be crowdsourced. Citizen scientists check each other's observations in a game, based on the photograph that was sent by the citizen scientist who estimated the water level with the app. This helps to either confirm or correct the original water level class estimate. Second, this game also trains new citizen scientists because they learn to read the virtual staff gauge while playing the game. When they later add a virtual staffgauge to their own photographs in the app, they tend to make fewer mistakes.

This thesis shows that it is possible for citizen scientists to collect water level class data with an app. In addition, a game enables the data to be quality controlled and for new citizen scientists to be trained. The crowdsourced data can be used to improve hydrological models that calculate when and how much water runs down the river.

PAPERS AND AUTHOR CONTRIBUTIONS

List of papers

This thesis is based on the following scientific publications:

I. Seibert, J., **B. Strobl**, S. Etter, P. Hummer, and H.J. van Meerveld (2019), Virtual staff gauges for crowd-based stream level observations, *Frontiers in Earth Science – Hydrosphere*, https://doi.org/10.3389/feart.2019.00070.

II. **Strobl, B.**, S. Etter, H.J. van Meerveld, and J. Seibert (2019), Accuracy of crowdsourced streamflow and stream level class estimates, *Hydrological Sciences Journal, Special Issue: Hydrological Data: Opportunities and Barriers*, https://doi.org/10.1080/02626667.2019.1578966.

III. **Strobl, B.**, S. Etter, H.J. van Meerveld, and J. Seibert (2019), The CrowdWater Game: a playful way to improve the accuracy of crowdsourced water level class data, *PLoS One*, https://doi.org/10.1371/journal.pone.0222579.

IV. **Strobl, B.**, S. Etter, H.J. van Meerveld, and J. Seibert (2020), Training citizen scientists through an online game developed for data quality control, *Geoscience Communication,* https://doi.org/10.5194/gc-3-109-2020.

Author contributions

Paper I: Jan Seibert and Ilja van Meerveld had the initial idea for the virtual staff gauge. The implementation in the smartphone application was discussed with all co-authors and tested in the field by Simon Etter and myself. Jan Seibert had the lead in writing the first draft of the manuscript, with all co-authors contributing to the writing and editing. Many of the figures were made by Simon Etter and myself.

Paper II: I designed this study together with all co-authors. The surveys were conducted by Simon Etter and myself. I analysed the data and the results were discussed with all co-authors. The first draft of the manuscript was written by myself with contributions by all co-authors. I created all the figures.

Paper III: I took the lead in designing and implementing the study, analysing the data, making the figures and writing the manuscript. Simon Etter helped with analysing the results for which expert judgement was needed. The study and results were discussed with all co-authors, who also contributed to the editing of the manuscript.

Paper IV: I took the lead in designing and implementing the training study, analysing the data, making the figures and writing the manuscript. Simon Etter helped with analysing the results for which expert judgement was needed. The study and the results were discussed with all co-authors, who also contributed to the editing of the manuscript.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Importance of hydrology and hydrological measurements

Hydrology is the study of *"water on and under the earth's surface"* [*Hendriks*, 2010, p. xi]. Hydrology seeks to fully understand the water cycle. It relates to how rivers flow, how they influence and are influenced by topography, and how groundwater moves and influences geological processes. Hydrology also studies water allocation, protection and rights issues [*Hornberger et al.*, 1998].

The field of hydrology goes back to antiquity, is very diverse and covers a lot of different sub-disciplines. Water is vital to life on earth but also represents severe risks, such as floods, droughts and water contamination. Therefore, information on the overall availability, quality and the temporal and spatial distribution of water on and under the earth's surface is crucial for societies and is the foundation for water management.

There are many types of water, such as atmospheric water, groundwater, soil moisture and surface water. In this thesis the focus is on surface water, which is *"water at the surface, whether stagnant in the form of surface storage or flowing in brooks or rivers, or as overland flow on slopes"* [*Hendriks*, 2010, p. 200]. More specifically, this thesis focuses on flowing surface water, i.e., streams and rivers.

There are many reasons why hydrological measurements are important. Hydrological data can help to improve our understanding of hydrological processes, to quantify our water resources and to check that the water quality is in compliance with regulations [*Western et al.*, 2005]. Streamflow data are used for river management, such as water allocation and for the calibration of hydrological models that can be used to help predict floods and droughts or climate change impacts [*Beven*, 2012].

There are different methods to collect surface water data, such as water levels and streamflow. Measuring water levels is generally easier, which can be done with a staff gauge, a water level recorder or a pressure transducer. Streamflow can be determined with volumetric gauging, in case of small flows. The velocity-area method is used for larger streams with a slow flow. Thereby, the river profile is surveyed and the flow velocity is measured using the float method, a current meter or an acoustic Doppler

current profiler. For mountainous rivers with turbulent flow, the salt dilution method is usually used to quantify the streamflow [*Hendriks*, 2010].

Because water levels are easier to quantify, usually gauging stations measure the water levels continuously and streamflow only occasionally. These measurements can then be combined via the rating curve (i.e., stage-discharge relationship), which requires several measurements and regular updates in case of river profile changes [*Hendriks*, 2010].

For many applications, streamflow data are needed. For example, most hydrological models require some streamflow data for calibration [*Beven*, 2012]. However, a study from Seibert and Vis [2016] indicates that water levels can also be used to calibrate a hydrological model. For humid catchments, the missing volumetric information is deduced from the annual precipitation amounts.

Hydrology is often limited by insufficient data, in particular in low-income regions [*Mulligan*, 2013; *Walker et al.*, 2016]. Such regions in particular, tend to have many water related issues, such as *"climatic water scarcity, high population-related demands, lack of – or poor – land and water management practices, poverty or significant inequalities in sharing water and its benefit"* [*Mulligan*, 2013, p. 750]. In addition, there is an overall decline in national hydrological and meteorological measurements networks [*Vörösmarty et al.*, 2001; *Fekete et al.*, 2012; *Ruhi et al.*, 2018]. Many new measurement approaches, such as remote sensing, geophysical methods and wireless sensor networks, have the potential to alleviate this situation. However, high spatiotemporal resolution streamflow measurements are still difficult to obtain. Citizen science therefore represents an innovative and novel approach to fill some of these gaps [*Buytaert et al.*, 2014].

1.2 Citizen science

1.2.1 Definition of citizen science

The Oxford English Dictionary defines citizen science as *"scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions"* [*Haklay*, 2014]. Thus, citizen science means that scientists and volunteers collaborate during some stage in the scientific process.

The common definition of a citizen scientist is *"a member of the general public who engages in scientific work, often in collaboration with or under the direction of professional scientists and scientific institutions; an amateur scientist"* [*Haklay*, 2014]. Another definition is *"a volunteer who collects and/or processes data as part of a scientific enquiry"* [*Silvertown*, 2009].

These definitions for "*citizen science*" and "*citizen scientist*" cover a wide range of projects and participants, yet the exact definition is still discussed in the citizen science community

[*Bonney et al.*, 2016; *Eitzel et al.*, 2017]. A common classification of the different degrees of involvement of citizen scientists was given by Haklay [2013]. Figure 1 shows the four different levels of involvement, according to this classification.

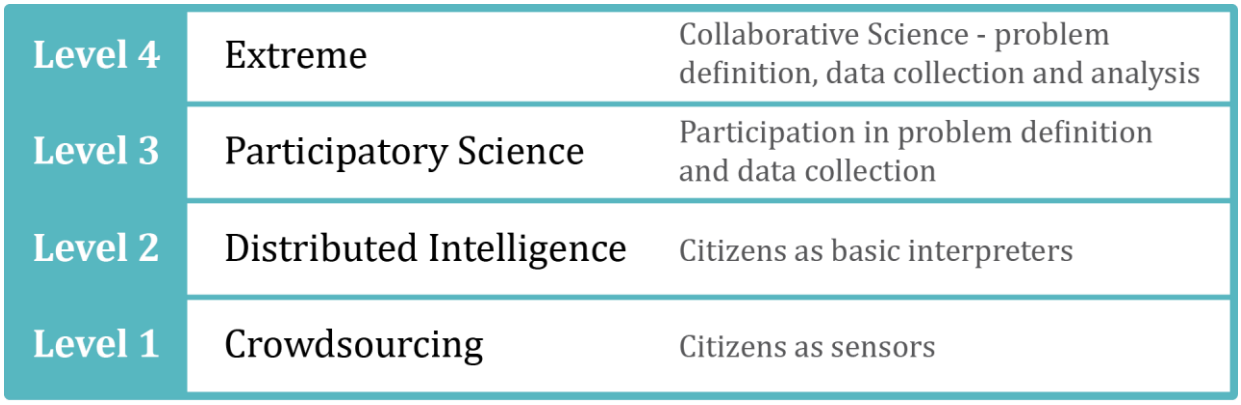


Figure 1: Levels of participation in citizen science. Figure adapted from Haklay [2013].

Most frequently, citizen scientists are engaged in projects with lower levels of participation, i.e., crowdsourcing and distributed intelligence. However, in some projects participants may have different tasks or degrees of involvement [*Haklay*, 2013]. An example for the lowest level of participation is volunteer computing [*Kloetzer et al.*, 2016]. Examples for projects that ask citizen scientists to interpret information are CoCoRaHS [*Reges et al.*, 2016], Galaxy Zoo [*Lintott et al.*, 2008], Foldit [*Cooper et al.*, 2010] and CrowdWater [*Paper I*]. An example that engages citizen scientists during the problem definition and data collection stage is for example the environmental justice case in Flint, Michigan [*Bellinger*, 2016; *Pieper et al.*, 2018]. Examples for extreme citizen science are anti-poaching or anti-illegal logging projects [*Matthias et al.*, 2014].

There are also other classification schemes available, such as subdividing projects into contributory, collaborative and co-created citizen science projects [*Bonney et al.*, 2009a]. Some example projects for this classification scheme are listed in Thornhill et al. [2019] or Bonney et al. [2009a].

1.2.2 History of citizen science

While the term *"citizen science"* is relatively new, the concept has been around for quite a long time [*Irwin*, 1995; *Bonney et al.*, 2009a]. Various projects are credited with being the first citizen science project, most commonly the Audubon Christmas Bird Count in the United States, which began collecting bird observations in 1990 [*Dunn et al.*, 2005]. Others mention diary entries of the cherry tree blossoming date in Kyoto, Japan in the 9th century [*Aono and Kazui*, 2008], lighthouse keepers collecting bird strike data starting in 1880, the National Weather Service Cooperative Observer Program starting in 1890 [*Bonney et al.*, 2009a] or volunteers recording the tide for two weeks in 1835 [*Cooper*, 2016].

Over the last decade the field of citizen science has grown enormously [*McKinley et al.*, 2017]. This recent advent can partly be accredited to mobile smartphones becoming

ubiquitous [*Silvertown*, 2009; *Graham et al.*, 2011; *Dickinson et al.*, 2012; *Paul et al.*, 2018]. Smartphones make it easy to immediately transfer any observation with a GPS-location and a timestamp. Furthermore, the option to add a photograph to any observation, facilitates quality control for project managers [*Dickinson et al.*, 2012]. Moreover, data collection via smartphones enables citizen scientists to have almost immediate access to their own data and other observations [*Graham et al.*, 2011] and to build online communities that can increase the motivation to participate [*Science Communication Unit - University of the West of England*, 2013].

The recent increase in citizen science projects is also manifested through the creation of the Citizen Science Association (www.citizenscience.org) in 2012 and the European Citizen Science Association (www.ecsa.citizen-science.net) in 2013. There are also national associations, e.g. in Switzerland called *"Citizen Science Schweiz"* that was launched in 2015 (www.schweiz-forscht.ch) or in Austria called *"Zentrum für Citizen Science"* also launched in 2015 (www.zentrumfuercitizenscience.at).

1.2.3 Citizen scientists: contribution patterns and motivations

The variability in the contributions from citizen scientists is similar in most citizen science projects: the majority of citizen scientists make one contribution and the vast majority of contributions are provided by a few dedicated citizen scientists [*Sauermann and Franzoni*, 2015]. FreshWater Watch reported that 1% of citizen scientists provided 47% of the observations [*August et al.*, 2019], 86% of participants who joined the CrowdHydrology project sent only one observation [*Lowry et al.*, 2019] and for the project iSpot more than half of all registered participants never uploaded any observation, but a few hundred of the 42,000 registered participants added *"hundreds or thousands"* of contributions each [*Silvertown et al.*, 2015].

In order to be able to retain citizen scientists and attract frequent contributors, it is important to analyse the motivations of citizen scientists to participate, which can vary significantly between projects. Project managers need to understand the relevant incentives in order to target the right audience and engage the public for longer [*Thornhill et al.*, 2019]. Motivations can differ between initially joining a project and sustained participation and can broadly be classified into intrinsic and extrinsic motivations [*West and Pateman*, 2016]. Factors inspiring citizen scientists to participate over a long period are feedback from the project, good communication between project organisers and citizen scientists, social interactions among citizen scientists and various reward systems [*West and Pateman*, 2016]. A study investigating the motivations of the participants of the CrowdWater project (see 3. CrowdWater) concluded that citizen scientists mostly decided to join in order to contribute to science, improve the wellbeing of society and to protect nature [*Etter et al.*, in review]. A survey investigating the motivations to join the CrowdWater game showed that citizen scientists enjoyed playing the game, were

interested in hydrology, wanted to be part of the CrowdWater community and enjoyed helping others [*Paper III*].

1.2.4 Benefits and limitations of citizen science

Citizen science projects have many advantages compared to more conventional research projects, yet these benefits depend considerably on the project and the research field. One of the main benefits of citizen science approaches is the possibility to gather vast amounts of data in a relatively short time [*Catlin-Groves*, 2012]. For example, Galaxy Zoo managed to collect 8 million galaxy classifications within 10 days of launching their site [*Clery*, 2011]. The project Phylo, a citizen science project for improving multiple genome sequence alignment, received more than 350,000 submitted solutions within their first year [*Kawrykow et al.*, 2012].

Citizen science also enables data collection with a unique spatial coverage [*Goodchild*, 2007]. For example, the Audubon Christmas Bird Count has over 70,000 annual participants and has collected data on 551 different bird species in the United States and southern Canada. This spatially distributed dataset enables researchers to investigate trends for the whole study region and also at smaller scales, e.g. at the level of states or provinces [*Soykan et al.*, 2016]. CoCoRaHS, a meteorological citizen science project, has approximately 20,000 active citizen scientists that provide data in the United States, Puerto Rico, the U.S. Virgin Islands and in 13 Canadian provinces [*Reges et al.*, 2016]. Another project with exceptional spatial coverage is the German project *"Mückenatlas"* (mosquito atlas), which asks participants to send in frozen mosquitos for identification. In this project more than 17,000 mosquitos including 39 species were collected [*Haklay*, 2015].

Citizen science projects sometimes take advantage of the unique local or individual knowledge of participants [*Wilson et al.*, 2018]. Examples for such projects are anti-illegal-logging projects [*Matthias et al.*, 2014] or a study about tropical resource monitoring [*Danielsen et al.*, 2014].

Citizen science can, furthermore, help to bridge the gap between science and society and introduce citizens to science [*Overdevest et al.*, 2004; *Bonney et al.*, 2009a; *Price and Lee*, 2013]. However, some researchers mention that education should not be the main aim of citizen science projects, as per definition, citizen science should primarily be a tool for scientific work [*Bonney et al.*, 2014]. Citizen scientists may learn field-specific information within the project [*Schrier*, 2017]. For example, Crall et al. [2013] found that by participating in a training programme, citizen scientists learned to better identify invasive plants. However, in some projects no learning effect took place [*Overdevest et al.*, 2004].

Some researchers perceive citizen science projects as outreach or educational projects [*Bonney et al.*, 2014]. This misconception can likely be reduced through further studies

that demonstrate the scientific validity of citizen science projects and data, by maintaining high scientific standards and by communicating the potential of such data and results.

While the potential of citizen science has been demonstrated by numerous successful projects worldwide and in many different scientific fields [*Cooper*, 2016], there are also some challenges. Citizen science is sometimes criticised for the unconventional data collection methods and often the data quality is doubted [*Catlin-Groves*, 2012; *Burgess et al.*, 2016; *Parrish et al.*, 2018; *Wilson et al.*, 2018; *Njue et al.*, 2019]. Although no generalised statements regarding the data quality can be made, many studies have demonstrated the accuracy and validity of their data (see 1.4.2 Data quality in citizen science).

Sensitive data, such as the participants' private information or the location of endangered species have to be considered carefully, consent has to be obtained and the data have to be protected whenever possible [*Haklay*, 2015]. These issues need to be handled on a project by project basis. Some environmental citizen science projects have chosen not to share the data publicly, in order to protect certain species [*Bowser and Wiggins*, 2015; *Ganzevoort et al.*, 2017; *de Vries et al.*, 2019] or to safeguard user privacy [*Newman et al.*, 2012; *Rey-Mazón et al.*, 2018].

Currently, there is a lack of diversity in citizen scientists. Haklay [2015] writes that the average citizen scientist is *"well educated, working in a job that provides enough income and working conditions for ample leisure, and with access to the internet as well as ownership of smartphones"*. Therefore, some projects struggle with a geographical bias, meaning that there are more projects in wealthy regions [*Buytaert et al.*, 2014; *Haklay et al.*, 2018]. Haklay et al. [2018] point out the irony of this, as many citizen science researchers talk about the potential of citizen science in less affluent regions [e.g. *Njue et al.*, 2019]. Some projects have nonetheless been successful in poorer regions, such as a water level project in Kenya [*Weeser et al.*, 2018] or SmartPhones4Water, a project for precipitation measurements in Nepal [*Davids et al.*, 2019]. Those projects sometimes had to adjust the incentives for participation. Buytaert et al. [2014] mention that citizen science projects in developing countries more commonly pay their citizen scientists for contributions. For citizen science to reach its full potential, strategies to be more inclusive will have to be explored. That innovative strategies can expand the user groups is shown in Extreme Citizen Science projects, where projects were designed to even include illiterate groups [*Haklay*, 2015].

1.3 Citizen science in hydrology

Citizen science in the field of hydrology is still a relatively new, but rapidly expanding field [*Buytaert et al.*, 2014; *Assumpção et al.*, 2018; *Njue et al.*, 2019]. Traditional hydrological measurements are collected with expensive equipment, which poses a challenge to citizen

science projects [*Buytaert et al.*, 2014]. Nonetheless, contributory citizen science projects are the most common form of participation [*Buytaert et al.*, 2014; *Njue et al.*, 2019].

A few hydrological citizen science projects aim at complementing the scarce streamflow data, typically by providing additional water level observations [*Buytaert et al.*, 2014; *Njue et al.*, 2019]. A successful hydrological citizen science project, called CrowdHydrology, partly inspired the CrowdWater project that is central to this thesis. CrowdHydrology collects water level data by placing a staff gauge in streams and rivers, mostly along hiking paths. A sign encourages people who happen to walk by to read the current water level and send a text message with this value. By 2019, the U.S.-based project had 120 stations, over 16,000 observations and over 8,000 participants [*Fienen and Lowry*, 2012; *Lowry and Fienen*, 2013; *Lowry et al.*, 2019]. A similar approach using fixed staff gauges in rivers was successfully used by a project in Kenya [*Rufino et al.*, 2018; *Weeser et al.*, 2018]. Both of these projects rely on actual water level gauges, which limits the number of possible locations. In addition to financial constraints, project organisers cannot visit an unlimited number of locations to place and maintain these staff gauges. In contrast, this thesis proposes a mobile smartphone application with a *virtual* staff gauge. This leads to a fully scalable approach because the citizen scientists place the virtual staff gauges in a photograph by themselves (see 3.2 The CrowdWater app). Thus, there is no extra cost per virtual staff gauge and no maintenance required. Unlike the physical staff gauge, the virtual staff gauge does not have absolute units. The accuracy of these data is explored in this thesis.

Hydrological citizen science projects extend beyond water level observations in rivers [*Buytaert et al.*, 2014]. For example, the project LOCSS (Lake Observations by Citizen Scientists and Satellites) uses water level gauges in lakes in order to obtain ground truth observations for satellites [*The University of North Carolina at Chapel Hill*, 2019]. Groundwater well water levels were monitored in a Canadian study [*Little et al.*, 2016]. Wet/ dry mapping of intermittent streams has been done in multiple citizen science projects [*Turner and Richter*, 2011; *Kampf et al.*, 2018; *Allen et al.*, 2019]. Other studies investigated qualitative observations of soil moisture, such as the *"boots and trousers"* method [*Rinderer et al.*, 2012, 2015]. In a different project, snow depth observations were collected by skiers and snowboarders who use an avalanche probe as a measuring stick [*Hill et al.*, 2018]. The Swedish meteorologist, Tor Bergeron, already successfully asked people to measure snow depth between 1941-1943 [*Bergeron*, 1949] and also measure rainfall [*Bergeron*, 1960] and to mail their observations via postcard. Lottig et al. [2014] used a long-term dataset (between 1938 and 2012) collected by citizen scientists, to study geographical differences and temporal trends in lake-water clarity in eight states in the United States.

Most citizen science projects in hydrology collect observations to assess water quality. For example, the HydroCrowd project [*Breuer et al.*, 2015] collected water samples to

investigate nitrogen concentrations in Germany with the help of students. A similar study in Canada examined various water quality parameters [*Jollymore et al.*, 2017]. The project FreshWater Watch crowdsources observations regarding algal presence, turbidity, water colour and nitrate and phosphate concentrations with the help of trained citizen scientists [*Castilla et al.*, 2015; *Thornhill et al.*, 2018]. A study in South, Central and North America asked citizen scientists to provide microscale information, such as point sources and bank vegetation conditions, in order to explain sub-basin variability of phosphate concentrations [*Loiselle et al.*, 2016]. Visible macro-plastic pollution can also be observed with the help of citizen scientists [*Emmerik and Schwarz*, 2020].

In addition to data collection, multiple investigations into hydrological modelling with citizen science data have been made to demonstrate the usefulness of such crowdsourced observations [*Assumpção et al.*, 2018; *Etter et al.*, 2018, 2020; *Mazzoleni et al.*, 2018; *Weeser et al.*, 2019]. For example, Weeser et al. [2019] showed that crowdsourced water level measurements could effectively calibrate a lumped hydrological model, in particular when including water balance or evapotranspiration data. Etter et al. [2020] showed that water level classes can be used to calibrate the HBV model even when the data have a low temporal resolution and contain some errors (which corresponds to a citizen science scenario).

1.4 Data quality

1.4.1 Data quality in hydrology

Hydrological data, just like any environmental data, are never fully accurate or complete. Whitfield [2012] writes that *"most environmental data suffer from missing observations, missing periods, and other forms of incompleteness"*. Even gauging station data can still contain errors [*McMillan et al.*, 2012; *Whitfield*, 2012; *Chao et al.*, 2015; *Kiang et al.*, 2018] due to errors in the stage and streamflow measurements, interpolation and extrapolation of the rating curve and changes in the stream cross-section [*McMillan et al.*, 2012]. According to McMillan et al. [2012] relative errors for streamflow vary depending on the streamflow volume and can be ±50–100% for low flows and ±10–20% for medium or high flows. Streamflow beyond the streambank is likely to have even higher errors. Westerberg et al. [2011] calculated rating curve related errors of −60% to +90% for low flows and ±20% for medium to high flows. Thus, data might have varying quality. It is therefore important that the *"fitness-for-purpose"* is considered. This means that data are only used for purposes for which the data quality is sufficient. Therefore, it is essential that the data quality is communicated and that data users consider such factors [*Whitfield*, 2012].

Even when including the error ranges mentioned above, gauged streamflow data likely have a higher data quality than crowdsourced data. However, while the temporal

resolution of gauging stations is high, they are more limited and less flexible with regard to spatial coverage than crowdsourced data, due to construction and maintenance costs [*Lowry et al.*, 2019]. Therefore, smaller catchments often remain ungauged [*Kirchner*, 2006; *Bishop et al.*, 2008]. Currently, observation networks are more likely to decrease the spatial resolution than to increase it [*Kundzewicz*, 1997; *Ruhi et al.*, 2018]. As an example, in the United States the number of stream gauges decreased by 21% between 1947 and 2016 [*Ruhi et al.*, 2018].

1.4.2 Data quality in citizen science

The issue of data quality receives a lot of attention in citizen science [*Engel and Voshell*, 2002; *Haklay*, 2010; *See et al.*, 2013; *Aceves-Bueno et al.*, 2017]. Many researchers and data users alike are worried about the data quality of less standardised and less well-established approaches [*Catlin-Groves*, 2012; *Burgess et al.*, 2016; *Parrish et al.*, 2018; *Wilson et al.*, 2018; *Njue et al.*, 2019]. Therefore, many citizen science projects begin by analysing the data quality for their innovative data collection strategies [*Turner and Richter*, 2011; *Rinderer et al.*, 2012, 2015; *Lowry and Fienen*, 2013; *Peckenham and Peckenham*, 2014; *Breuer et al.*, 2015; *Le Coz et al.*, 2016; *Little et al.*, 2016; *Weeser et al.*, 2018]. Many of these projects found promising results regarding the accuracy of the crowdsourced data. For example, Lowry and Fienen [2013] analysed water level data that were collected by hikers by reading a staff gauge. The results showed that the crowdsourced data were almost as good as data from a pressure transducer. A similar approach to collect water levels was chosen by Weeser et al. [2018], who also found that the crowdsourced data were comparable to those from data loggers. Groundwater levels were collected and analysed in a study by Little et al. [2016], who found that the absolute difference of the well readings of 2 to 11 mm was sufficiently small and that crowdsourcing therefore provided a cheap and effective method for water resources management.

Citizen science projects try several approaches to maintain a high data quality, often depending on the field and level of engagement. These approaches are categorised as prevention (before data collection) or correction (after data collection). Examples for methods before data collection are: intuitive data collection approaches, standardised protocols and training new citizen scientists. Examples for methods after data collection are: automatic data filtering, collecting many observations that can be cross-checked for consistency and quality control either through experts or through citizen scientists [Wiggins et al., 2011].

There are many projects that try to develop intuitive and simple data collection approaches that are easy to implement and therefore avoid mistakes. Examples of such approaches are a qualitative soil moisture scale adapted to different regions [*Rinderer et al.*, 2012, 2015], wet/ dry mapping of intermittent streams [*Turner and Richter*, 2011; *Kampf et al.*, 2018] and the virtual staff gauge for water level class estimates [*Paper I*].

Some projects opt for strictly standardised data collection protocols to minimise misunderstandings that enable citizen scientists to follow clear instructions. An example is a snow observation project in the United States [*Dickerson-Lange et al.*, 2016] or a bird survey across three citizen science projects [*Jones et al.*, 2018]. Many projects have training material or training days, in order to ensure that citizen scientists have a basic understanding of the topic and the applied methodology. FreshWater Watch provides a one-day training to volunteers [*Thornhill et al.*, 2019] and CoCoRaHS has developed training animations [*Reges et al.*, 2016].

There are many methods of data quality control after data collection. A study investigating hail reports automatically filters reports according to the meteorological condition [*Barras et al.*, 2019]. Filtering can only be done in projects where likely limits on values can be pre-determined, which is not feasible in many projects. In a study investigating crowdsourced hydro-meteorological data, researchers benefited from collecting many observations as they were able to cross-check these observations with neighbouring observations for consistency [*Walker et al.*, 2016]. For this approach, data with close proximity and collected at a similar time are needed, which is also not always possible. In smaller projects, data quality control can be done with experts [*Wiggins et al.*, 2011]. Alternatively, quality control can also be crowdsourced, as for example in the projects Pattern Perception [*Koch and Stisen*, 2017] or Cyclone center [*Hennon et al.*, 2015] that ask citizen scientists to visually compare spatial patterns, or the projects Snapshot Serengeti [*Swanson et al.*, 2016] or Cropland Capture [*See et al.*, 2014] that require citizen scientists to assess photographs. Another example for crowdsourced quality control is the CrowdWater game (see 3.3 The CrowdWater game).

The fundamental questions regarding data quality that need to be answered by any new citizen science project are:

1. Can citizen scientists collect data with the required quality? And if not, how can the data quality be improved either before or after data collection?
2. Is the quality of the resulting data good enough for the intended purpose?

Both of these questions were addressed for the CrowdWater project (see 3. CrowdWater). The first question is discussed in this thesis, as well as in Papers I-IV. Section 8. Outlook briefly touches upon the second question, which is further discussed in a related thesis by Simon Etter, as well as in other publications [*Seibert and Vis*, 2016; *van Meerveld et al.*, 2017; *Etter et al.*, 2018, 2020].

2

SCOPE OF THE THESIS

Crowdsourced hydrological data might enhance the spatial coverage of hydrological measurements. Therefore, the overall aim of this thesis is to investigate the quality and potential of crowdsourced hydrological data, in particular water level class observations collected with a virtual staff gauge. This thesis examines the general feasibility of this approach, characterises the quality of the collected data, and investigates how the data quality can be further improved.

The research questions of this thesis were:

1. Is the virtual staff gauge concept suitable to crowdsource water level data?
    a. What errors occur when citizen scientists use the virtual staff gauge?
    b. How accurately can citizen scientists collect water level class data with the virtual staff gauge?
2. Is it possible to crowdsource data quality control?
3. Can a game for data quality control also be used to train new citizen scientists?

To answer the first research question, the benefits, limitations and errors of the virtual staff gauge approach are discussed [*Paper I*]. This virtual staff gauge approach is also compared with other crowdsourcing methods for hydrological data, by comparing the accuracy of water level class and streamflow estimates [*Paper II*]. For the second research question, a gamified method to crowdsource data quality control to further improve the data quality is investigated [*Paper III*]. The use of the game for training new citizen scientists is explored to answer the third research question [*Paper IV*]. The connections between these four papers are illustrated in Figure 2. In the outlook, preliminary modelling results are presented that show an example of how quality-controlled crowdsourced water level class observations can be used for hydrological model calibration in the future.

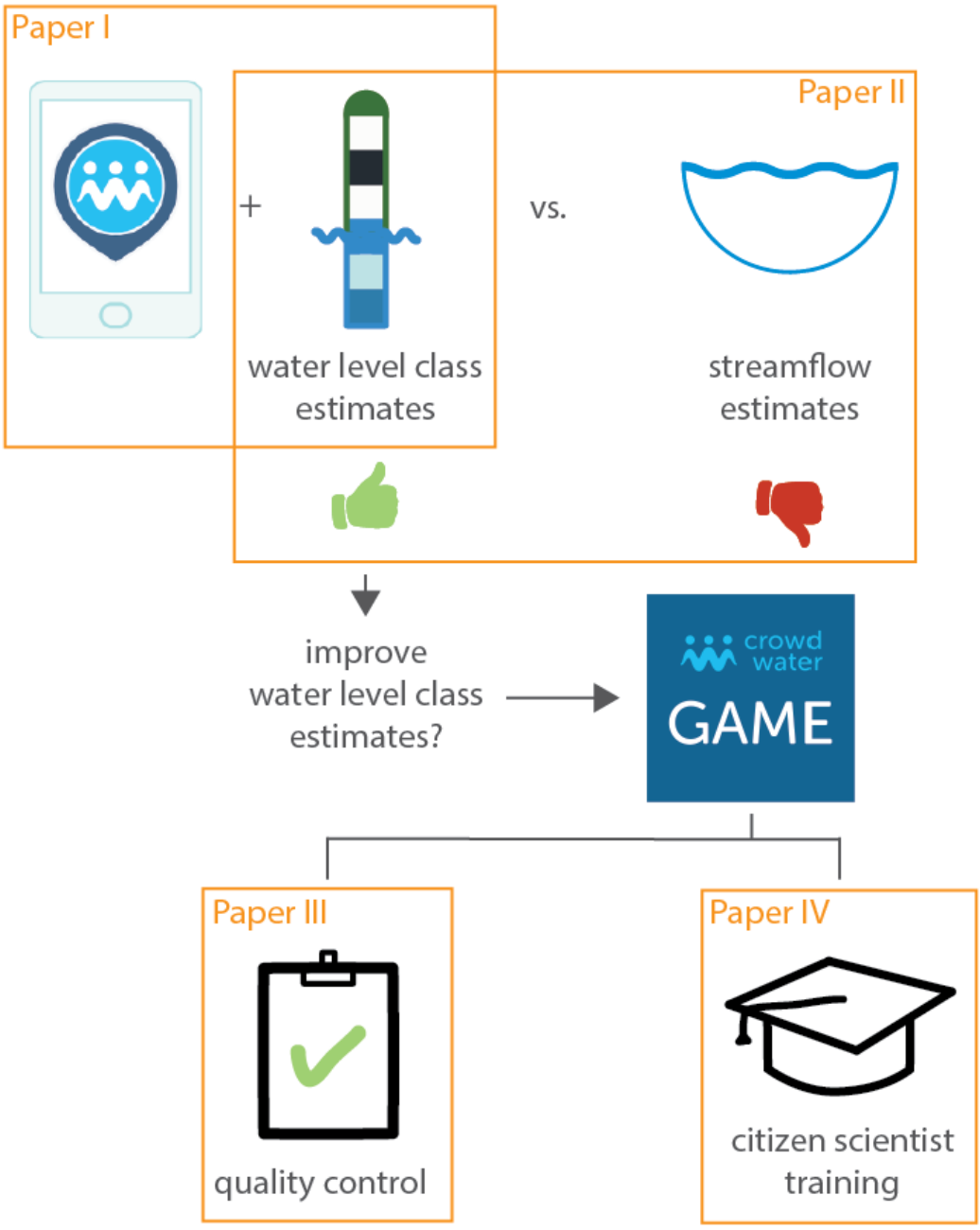


Figure 2: Overview of the thesis and themes of the papers.

3

CROWDWATER

3.1 The CrowdWater project

The research questions of this thesis were investigated within the CrowdWater project. CrowdWater is a hydrological citizen science project (www.crowdwater.ch), where any citizen can participate in the hydrological data collection. This is done through a mobile smartphone application, which is used for data collection, and an online game, which is used for quality control. The CrowdWater app and the data are freely available.

The project was launched at the Department of Geography, University of Zurich by Prof. Dr. Jan Seibert and Dr. Ilja van Meerveld in spring 2016 and was funded by the Swiss National Science Foundation. Between 2016 and 2020, two Ph.D. students (Simon Etter and myself) worked on this project.

Citizen scientists were recruited through various methods, such as public talks, science fairs, social media, student events, conferences, press releases and personal invitations via the project members' networks. To help with the outreach and recruitment, we produced or commissioned tutorial and motivational videos, graphics and a mascot called Droppy (Figure 3).

According to Haklay's levels of citizen science [*Haklay*, 2013], CrowdWater fits into the level of citizen engagement called *"Distributed Intelligence"*. CrowdWater citizen scientists collect data and provide some basic interpretation as well, such as estimating water levels (see 3.2 The CrowdWater app). In the CrowdWater game players interpret water levels by comparing photographs, which is also considered *"Distributed Intelligence"*. According to a different classification scheme by Bonney et al. [2009b], CrowdWater can also be considered a *"contributory"* citizen science project.

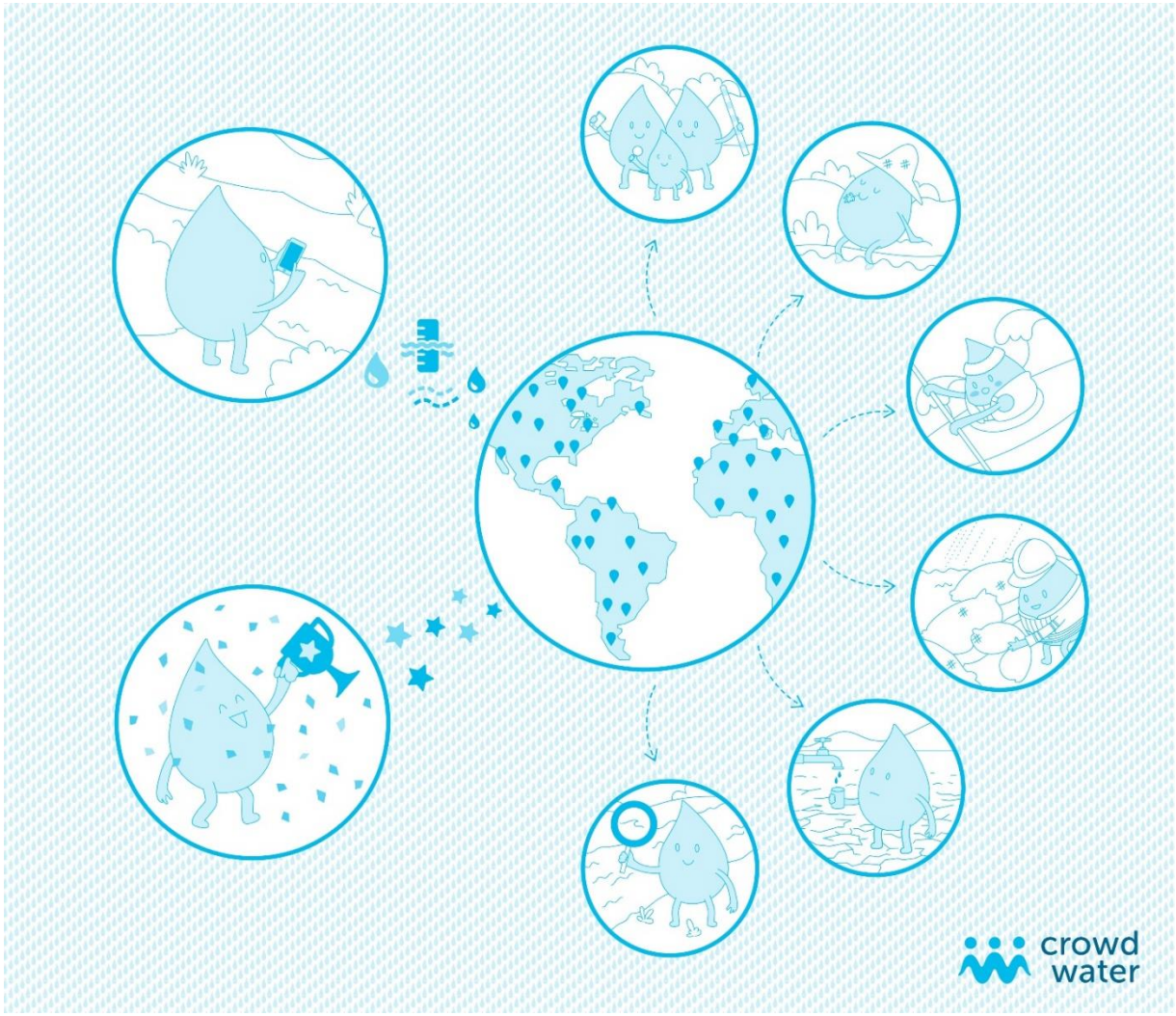


Figure 3: An example of an explanatory graphic produced by the University of Zurich's Multimedia & E-Learning-Services. The graphic explains the concept of the CrowdWater project as well as several possible data uses by citizen scientists. The infographic is centred around the CrowdWater mascot: Droppy. Figure by Tara von Grebel.

3.2 The CrowdWater app

The CrowdWater smartphone application (from here on called app) can be used to collect observations or estimates of water level classes, soil moisture, the state of temporary streams and the amount of plastic in streams. The app was first launched in spring 2017. By December 2, 2019 the CrowdWater app had 663 users and included 10,157 observations at 2616 different locations worldwide. This thesis focuses on the water level class estimates, for which 4948 observations at 931 locations worldwide had been submitted by December 2, 2019. The 10 citizen scientists who submitted the most water level class observations, contributed 59% of all water level class observations. The CrowdWater app was co-designed and produced by SPOTTERON (www.spotteron.net), an Austrian-based company specialising in citizen science smartphone applications.

Citizen scientists start collecting water level class time series by taking a reference picture, i.e., a photograph of a river to which they add a sticker-like virtual staff gauge with

10 classes (Figure 4). For follow-up observations, they compare the current water level with the original reference picture and submit the current water level class, as well as a new photograph of the stream [*Paper I, Seibert et al.*, 2019].

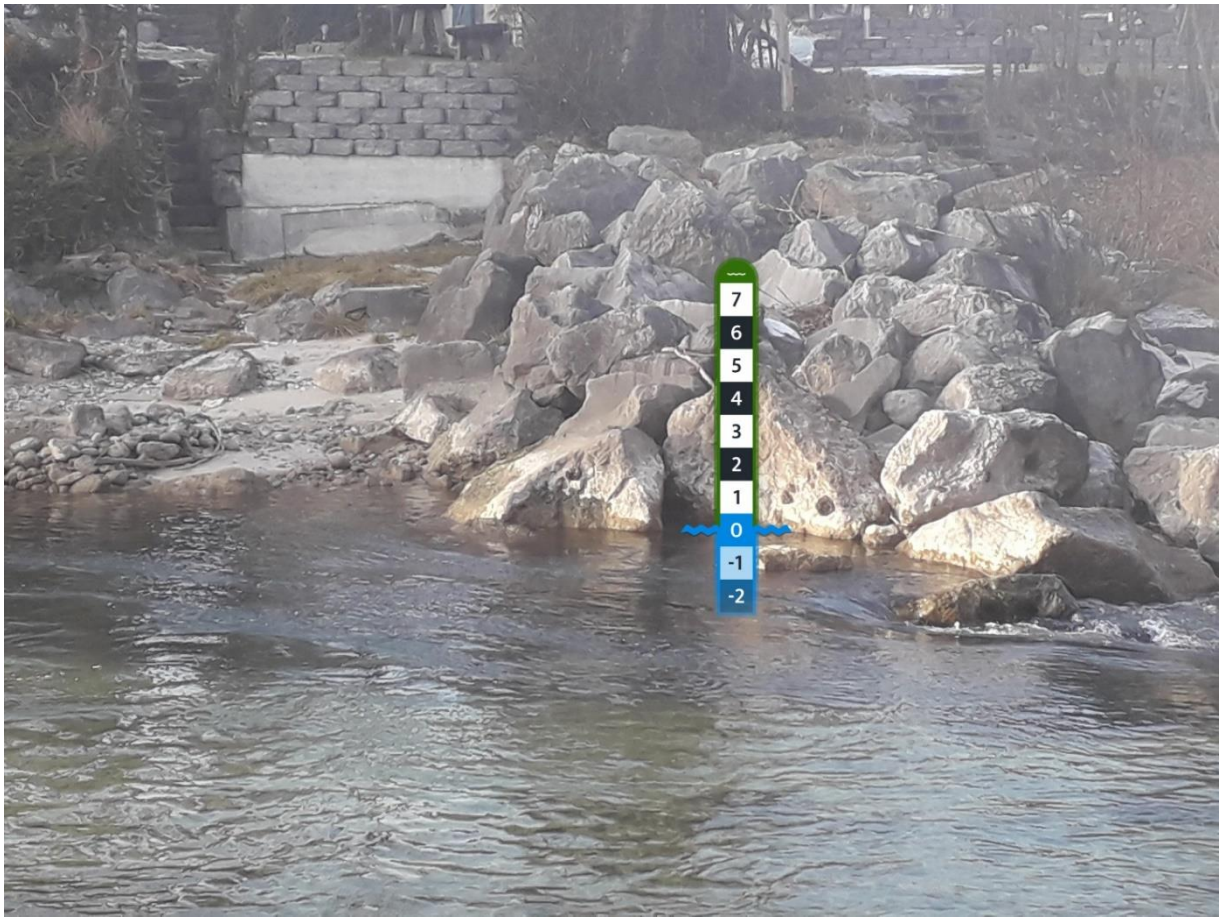


Figure 4: A reference picture is a photo with a sticker-like virtual staff gauge added to it.

The CrowdWater app also includes features that are similar to social media. Citizen scientists can like and comment on each other's observations and can follow each other's contributions. These features were included in the app to establish a community spirit, which can promote long-term engagement with a project [*Jennett et al.*, 2016]. These social media features are used by some of the citizen scientists, but many citizen scientists do not use them. Unfortunately, no data are available to quantify how many citizen scientists are actively using these features. It is possible that it will simply take some time until these features are more widely adopted by the community. For further information regarding the CrowdWater mobile app, see Papers I, III and IV.

3.3 The CrowdWater game

The CrowdWater game is a web-based citizen science game to check and improve the accuracy of the water level class data. It was launched in cooperation with SPOTTERON in spring 2018. The game enables multiple citizen scientists to check the water level class observations submitted via the CrowdWater app, thereby improving the data quality. The

game only controls the data quality of water level classes (and not the other three observation types collected through the app), because the water level is clearly visible on a photograph and can thus be assessed without having to be next to the stream.

In this web-based game, players compare the reference picture with the virtual staff gauge to a follow-up observation and vote on a water level class (Figure 5). Per observation 15-50 votes are collected to determine the mean water level class, called the mean vote.

Each game round contains two types of picture pairs: unclassified and classified. For unclassified picture pairs, the correct water level class value is still unknown because fewer than 15 players have voted on them so far. Classified picture pairs have been voted on by at least 15 players, therefore the mean of all votes already yields a reliable water level class value that is assumed to be correct.

Players can also report a picture pair if for example they are unable to see the water level on the picture (e.g. the picture is too dark) or for technical issues (e.g. the staff gauge is missing). The reasons for the report can be stated during the submission.

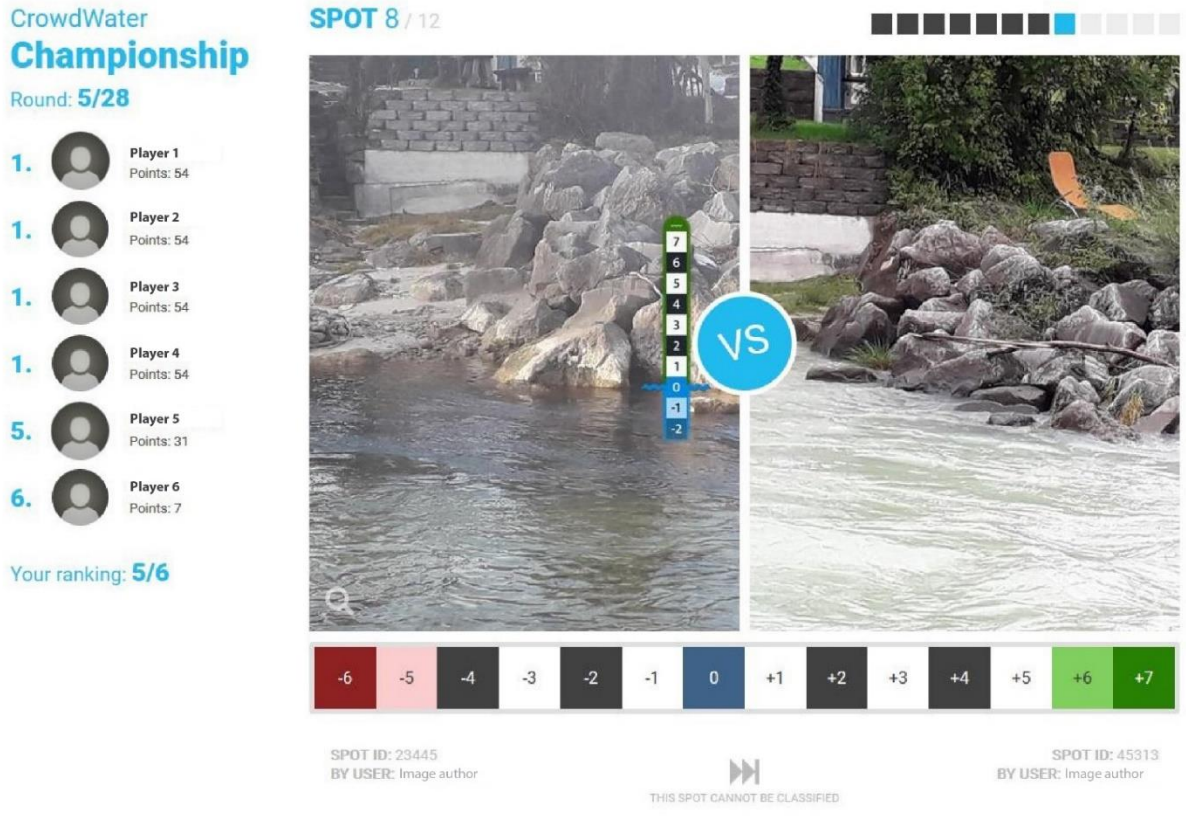


Figure 5: Screenshot of the CrowdWater game. The left picture shows the reference picture with the virtual staff gauge. The right picture shows the stream at a later date (an update), for which the player estimates the water level class. The water level class can be selected on the horizontal gauge classes below the pictures. Left of the pictures, the top ranked players of this round are shown (anonymised in this illustration). Figure taken from Paper III.

The CrowdWater game is a casual citizen science game, which means that little time, prior knowledge, experience or training is needed to start playing the game [*Crowston and Prestopnik*, 2013]. The game consists of daily rounds and monthly championships. Points

(6 points for a correct vote on a classified picture pair, 4 points for an error of one class and 0 points for a larger error, 3 points for unclassified picture pairs or reports) can be gained for each picture pair, which are summed up during each round and championship.

The CrowdWater game automatically receives the CrowdWater app water level class pictures, hence the number of possible picture pairings is continuously growing. By December 2, 2019, the CrowdWater game had 3964 picture pairs (1599 classified and 2365 unclassified picture pairs) and 209 players.

The CrowdWater game forms the basis for Papers III and IV. The focus of Paper III is on crowdsourcing quality control through a gamified approach, whereas the focus of Paper IV is on training new citizen scientists through the game. For further information regarding the CrowdWater game, see Papers III and IV.

4

METHODS AND DATA

4.1 CrowdWater app and game data

Data collected in connection with the CrowdWater project (see 3. CrowdWater), including the CrowdWater app, the CrowdWater game, as well as additional surveys, form the basis for the analysis of this thesis. The CrowdWater app and CrowdWater game datasets used are subsets of the current datasets, based on the data that were available at the time of the publications (Table 1). The CrowdWater project and crowdsourcing efforts are ongoing. Therefore, more data are currently available.

Table 1: Summary of datasets used in this thesis. The time period of the subsets of the CrowdWater app and CrowdWater game were determined by the publication of the respective papers.

Data source	Time period of subset	Chapter	Paper
CrowdWater app	Spring 2017 – September 3, 2018	4.2	I
Survey		4.3	II
CrowdWater game	Spring 2018 - February 28, 2019	4.4	III
Survey		4.5	IV

The CrowdWater app data are publicly available and can be freely downloaded from the CrowdWater homepage (www.crowdwater.ch). Each observation also includes the location, the timestamp and an identifier of the citizen scientist who made the observation (User-ID). Usernames and e-mail addresses are excluded from the public download to protect the privacy of all citizen scientists. The e-mail address is only stored and used for administrative purposes and is not publicly visible. Citizen scientists and their data are protected by the General Data Protection Regulation established by the European Union.

A citizen scientist can choose any username and does not need to provide his or her real name. Demographic data, such as age, education or profession, can be added to the user profile. However, very few citizen scientists choose to do so. Therefore, we do not know

the demographic data of the CrowdWater citizen scientists, either from the app or the game.

All citizen scientists can see all observations, not just their own. However, they are only able to edit their own observations. Administrators and appointed moderators can edit observations from other citizen scientists. This is intended for quality control and allows erroneous observations to be removed or corrected. Reference pictures that were checked and deemed suitable by an administrator can be *"locked"* within the dataset. This shows citizen scientists and data users that the reference picture is suitable for further observations. Citizen scientists can no longer edit a locked reference picture, to ensure that all follow-up observations refer to the same reference picture.

There are no strict guidelines for when a reference picture with a badly placed virtual staff gauge is removed from the dataset. In general, reference pictures are removed when they do not enable further observations at this location, e.g., when the class zero of the virtual staff gauge is not located on the water surface. The removal of a reference picture is admittedly somewhat subjective and depends on the administrator. Location errors are only noticeable if they are severe, e.g., the location is supposedly next to a large river, but there is no river on the map. The locations of smaller rivers are harder to verify, as they may not be shown on the map. Currently, the backend of the CrowdWater app and dataset does not log the changes that are made to observations, nor who has made the changes (i.e., an administrator or the citizen scientist). Such an addition would enable better traceability and improve the metadata and, therefore, should be implemented in the future.

The CrowdWater game data are not yet publicly available but may be merged with the CrowdWater app data in the future. User-IDs, as well as location and observation IDs for the CrowdWater app and game, are identical, which facilitates merging these datasets.

4.2 Issues related to the placement of the virtual staff gauge

This section focuses on data submitted through the CrowdWater app and in particular on the reference pictures. It quantifies the errors that occurred when citizen scientists placed the virtual staff gauge onto river photographs. Reducing these errors by training new citizen scientists is discussed in a later chapter (see 4.5 Training citizen scientists).

For the analysis of the CrowdWater app in Paper I, all app submissions collected between the first launch (spring 2017) and September 3, 2018 were included. At this point 2431 observations had been submitted by 218 citizen scientists, including roughly 500 water level class reference pictures, i.e., pictures with a virtual staff gauge.

Based on this dataset, we noted that some reference pictures had issues with the placement of the virtual staff gauge. Therefore, we analysed and quantified the frequencies of different error categories (staff gauge size problem, staff gauge placement

problem or unsuitable location). An exact quantification of the frequencies was not feasible, as not all changes in the dataset could be traced anymore because either an administrator or citizen scientist could have deleted an unsuitable spot. However, rough estimates were still possible.

4.3 Accuracy of water level class and streamflow estimates

We assessed people's abilities to estimate hydrological data with surveys (see Paper II). Two different types of estimates were assessed: estimates of streamflow based on individual streamflow factors (width, mean depth and flow velocity; Q_{factor}) and estimates of the water level class with the virtual staff gauge approach (Q_{level}).

We conducted 16 surveys at 10 different rivers or streams in Switzerland. A total of 517 passers-by participated. The 10 rivers were grouped into different size categories based on the mean annual streamflow: XS: ≤1 m^3/s (three streams), S: >1–50 m^3/s (five rivers), M: >50–200 m^3/s (one river), and L: >200 m^3/s (one river). We had already set the virtual staff gauge in a photo during a previous visit of the location, so that participants could compare the current water level with a potentially different water level in the reference picture.

We calculated streamflow estimates (Q_{factor} and Q_{level}) into a relative streamflow estimate, i.e., the streamflow estimate divided by the measured streamflow value times 100%, so that streamflow estimates for different streams could be compared. A value of 100% represents a perfect estimate, smaller values indicate an underestimation and larger values an overestimation. Estimates under 50% or over 150% are considered outliers in this analysis. Please note that the definition of outliers is not entirely consistent throughout this thesis, as the analyses differ between the publications. For the water level class estimates, we determined how many classes the estimate was off from our estimate of the water level class (which we assume to be correct).

In order to be able to compare the water level class estimates collected via the survey at Swiss rivers with the streamflow estimates, water level classes were re-calculated to a corresponding streamflow value (m^3/s). For stream locations with a nearby official gauging station, the classes of the virtual staff gauge were converted to streamflow classes by measuring the stream depth that corresponded to each water level class (midpoint and boundaries) and then using the available rating curve to find the corresponding streamflow value. For stream locations without an official gauging station, additional stream profile measurements were taken in order to calculate the corresponding streamflow with the Manning-Strickler formula [*Manning*, 1891].

4.4 Crowdsourced data quality control

In Paper III, the suitability of the CrowdWater game for crowdsourced data quality control is analysed. The analysis of the CrowdWater game dataset was finished on February 28, 2019, with 2326 picture pairs in the game (846 classified (≥ 15 game votes) and 1480 unclassified picture pairs). In total, 153 players contributed at least one vote in the game, but only 58 players played more than two full rounds.

The mean game vote and the original water level class submitted through the app were compared, in order to assess if the CrowdWater game is a suitable mechanism to correct erroneous app submissions. For all classified picture pairs (i.e., at least 15 votes), we calculated the mean game vote between the 10^{th} and the 90^{th} percentile (to exclude outlier votes). The comparison between the mean game vote and the app value can lead to one of three outcomes:

1. No water level class correction is needed, i.e., both values state the same water level class and the original app value can be confirmed.
2. A correction of a water level class is needed, i.e., the two values differ and either of them needs to be corrected.
3. The calculation of the mean game vote yields a higher water level class resolution (e.g. if half of all players vote for class one and the other half vote for class two, the resulting mean vote is 1.5). In this case, the original app value of one was not incorrect, but the mean game vote nonetheless delivers a higher resolution and therefore contains more information.

For all observations that required a correction, the pictures were assessed through expert judgement by the project members (Simon Etter and myself). The possible outcomes of this assessment were as follows:

- The original app value was better.
- The mean game vote was better.
- The correct value was between the original app value and the mean game vote.
- The observation should have been reported through the report function, as voting on a water level class was not possible.

We calculated a mean accuracy per player (the absolute difference between the player's vote and the mean game vote, averaged over all their votes and subtracted from 10) and plotted as a function of playing frequency. We divided the players between novice and regular players and tested the difference between these groups with the Mann-Whitney test ($p < 0.05$).

4.5 Training citizen scientists

For Paper IV, we conducted a computer-based training study with 52 participants to assess if citizen scientists can be trained by playing the CrowdWater game. We specifically investigated if participants improved the placement of the virtual staff gauge after training (playing the CrowdWater game for 50 picture pairs).

The participants' performance in placing the virtual staff gauge was quantified with a placement score. This placement score is the sum of score sub-categories, such as the selection of a good stream picture, the choice of virtual staff gauge and the angle and positioning of the virtual staff gauge within the stream picture. The resulting placement score ranges from 0 to 13. A score of 13 means that all sub-categories had full marks, though a score of 10 or higher is considered a good placement score.

The participants' performance during the training was quantified with the game score. Participants voted on 50 picture pairs, each with a maximum of 6 points. Therefore, a score up to 300 points could be gained but a score of 245 or higher was still considered a good game score.

To assess the differences of the placement score before and after the training, we analysed the training dataset with the Wilcoxon test ($p < 0.05$). Many participants already performed well before the training. We therefore also examined the subset of participants that had a low placement score (< 10 points) before the training, as we expected larger improvements through the training.

5 RESULTS

5.1 Virtual staff gauge

The virtual staff gauge enables citizen scientists to submit water level class observations. This has been demonstrated by CrowdWater app users since the launch of the app. Over the project's duration, several locations have received numerous updates (up to 582 observations at the same location, 14 locations with more than 51 observations, and 56 locations with more than 11 observations) and a time series of water level classes, reflecting the rivers' dynamics, could be established (Figure 6). Most of these locations tend to be visited by the same citizen scientist on a regular basis, some with an almost daily frequency. The location with the most water level class updates so far, is at the Königseeache river in Austria. This time series consists of 582 observations between December 2017 and December 2019, equalling a contribution frequency of a new observation every 1.2 days (Figure 7). Overall 394 citizen scientists have made at least one water level class observation and the median number of water level class observations per citizen scientist is two. However, 50 citizen scientist have made more than ten observations.

Figure 6: Reference pictures of the top 10 locations in terms of number of updates made by December 2, 2019. The locations are sorted according to number of updates.

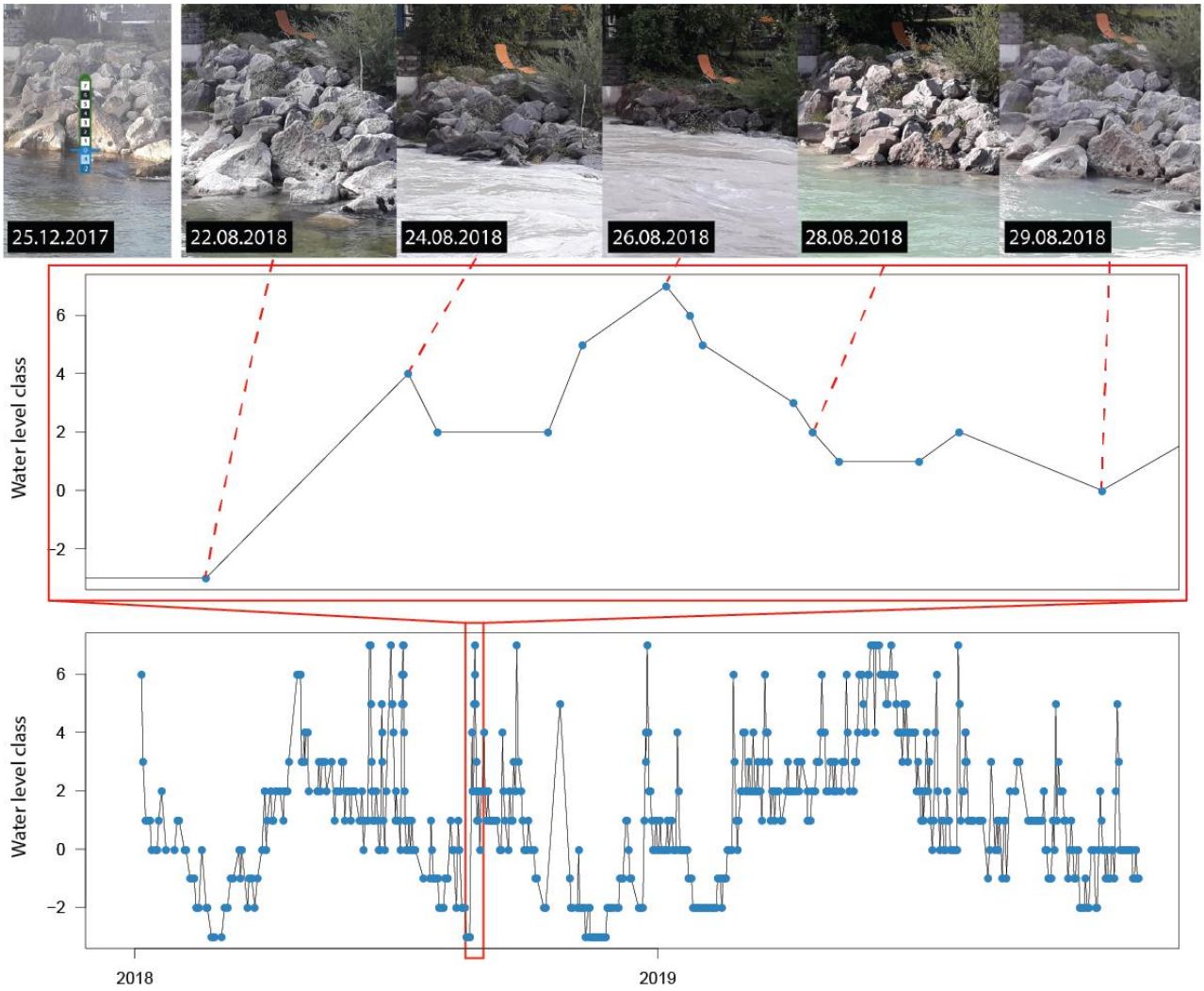


Figure 7: Time series of water level class observations at the Königseeache river in Austria. This time series demonstrates the potential of such crowdsourced observations. River dynamics are clearly visible, even though water levels are collected in classes instead of metric values (bottom plot). As an example a subset of the time series is highlighted (middle plot) and illustrated with the pictures submitted through the CrowdWater app (top plot). Figure adapted from Paper I.

Some submissions through the CrowdWater app, in particular water level class reference pictures, contained mistakes. We estimate that this occurred in about 10% of reference pictures. The most common mistakes were staff gauge scaling problems, staff gauge placement problems and the use of an unsuitable location (Table 2). These mistakes can lead to difficulties when trying to update the water level class for these locations. Therefore, they were usually excluded from the dataset and Paper IV analysed if training could reduce these errors. Based on the survey presented in Paper II, about half of first time users (48%) estimated the correct water level class and 40% were only one class off, indicating that a majority of first time users understood the virtual staff gauge concept. While these mistakes are expected to be lower for regular CrowdWater app users, they are lowered even further through crowdsourced quality control, as shown in Paper III.

Table 2: Overview of errors made by app-users grouped into broader error categories and frequency of occurrence. +++: occasional = more than 10 times; ++: seldom = 5-10 times; +: rare = less than 5 times; / = not quantifiable. Based on the 500 reference pictures available at the time of the study. Table taken from Paper I.

Error type		Frequency of occurrence
Staff gauge size problem	Staff gauge too big	+++
	Staff gauge too small	+
Staff gauge placement problem	Wrong angle	+++
	Staff gauge not on the water surface	+++
Unsuitable location	Lack of reference structure for water level identification	++
	Structure hidden by vegetation or snow	+
	Unclear which structure to use	+
	River bank too far away	++
	Poor image quality	+
	Site not easily accessible	/
	No suitable site for staff gauge placement available	/
	Changes in the rating curve	+
	Multiple measurement sites at (almost) the same location	+
	Testing (e.g. beer glasses, not a river, out of train, etc.)	++

5.2 Accuracy of water level class and streamflow estimates

The accuracy of the hydrological estimates received during our surveys at Swiss rivers differed significantly between the streamflow calculated out of streamflow factors (width, mean depth and flow velocity combined into Q_{factor}) and streamflow calculated out of the water level class with the virtual staff gauge approach (Q_{level}). The Q_{level} estimates were

more accurate and had fewer outliers than the Q_{factor} estimates (Table 3 and Figure 8), except for small streams for which the Q_{factor} had a smaller interquartile range. This anomaly was influenced by two specific survey locations. For further information please see Paper II. Whilst the Q_{level} overall was more accurate than the Q_{factor}, estimates still contained errors and sometimes were under- or overestimated. Therefore, methods that reduce these errors were explored in Paper III (see 5.3 Crowdsourced data quality control).

Table 3: Interquartile range, under- and overestimations for streamflow estimates calculated based on the streamflow factors (width, mean depth and flow velocity; Q_{factor}) and streamflow calculated based on the water level class (Q_{level}).

	Q_{factor}	Q_{level}
Interquartile range	30-163%	67-157%
Under- or overestimations (percentage of estimates < 50% or > 150%)	66%	39%
Extreme under- or overestimations (percentage of estimates < 10% or > 1000%)	11%	3%

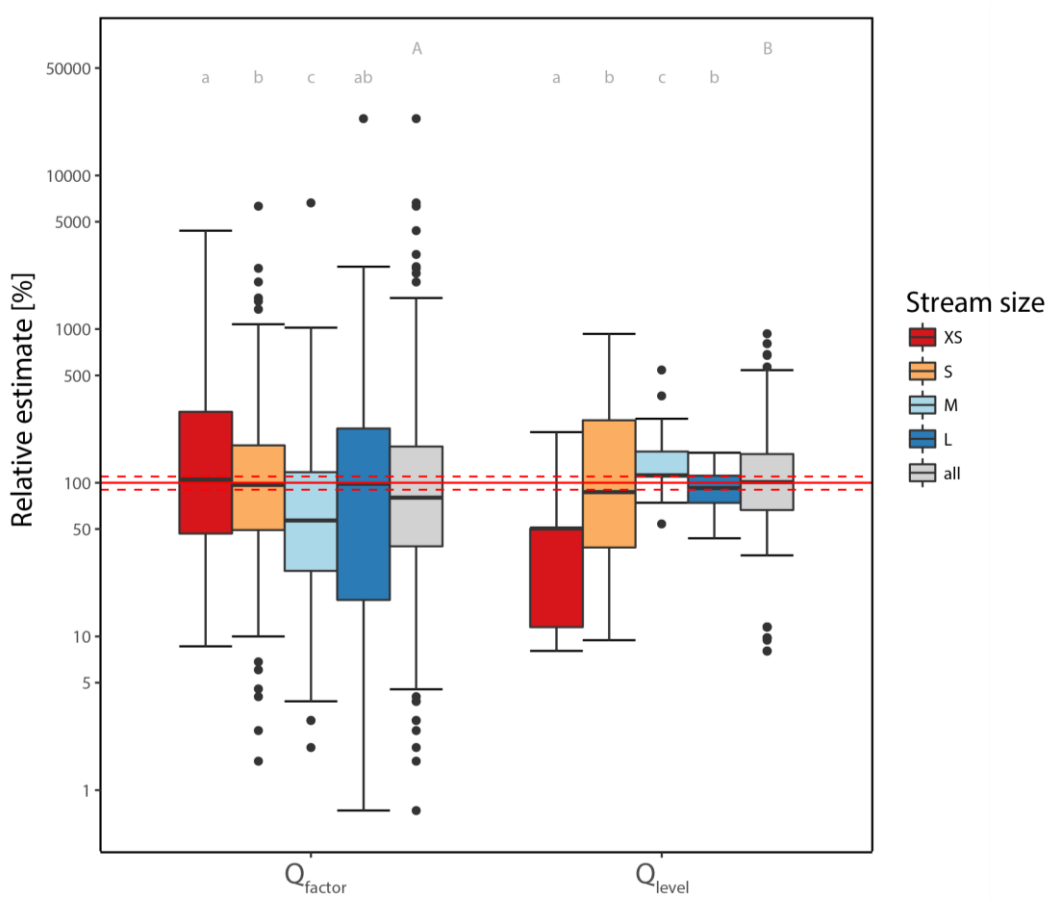


Figure 8: Boxplot of the relative estimates of Q_{factor} and Q_{level} for each stream size class and all surveys combined. The statistically significantly different medians are indicated by different upper case letters (combined data from all surveys) and different lower case letters (per stream size class). The solid (red) line at 100% indicates that the estimate is the same as the measured value; the dashed (red) lines indicate the 10% uncertainty band for the measured streamflow. The box indicates values between the interquartile range and points indicate values more than 1.5 times the interquartile range above or below the 25th or 75th quantile. Figure adapted from Paper II.

5.3 Crowdsourced data quality control

The CrowdWater game proved to be a successful method to crowdsource data quality control. The mean game vote and the original app submission agreed about the water level class (with a maximum difference of half a class) for 70% of the picture pairs. In 30% (252) of picture pairs the mean game vote and original app submission differed by at least one class. These picture pairs were evaluated through expert judgement (Simon Etter and myself). For 74% of the picture pairs for which the app and mean game vote water level class differed, the mean game vote was better than the original app submission and in 9% of picture pairs the original app submission was better. For 8% of the picture pairs, the correct value was in between the mean game vote and the original app submission. The CrowdWater game has a built-in reporting function, which enables players to report picture pairs, e.g. if there is an issue with the pictures and the water level class cannot be voted for. For 9% of the picture pairs, the players should have reported the pictures through this report function.

Figure 9 shows all classified observations (observations with ≥ 15 game votes). The mean game vote (red triangle) has a higher water level class resolution than the original app submission (orange star) and, therefore, also indicates if a water level is likely in between two classes. The figure also shows a relatively large agreement in observations that have a similar water level to the reference picture, i.e., that are in water level class zero.

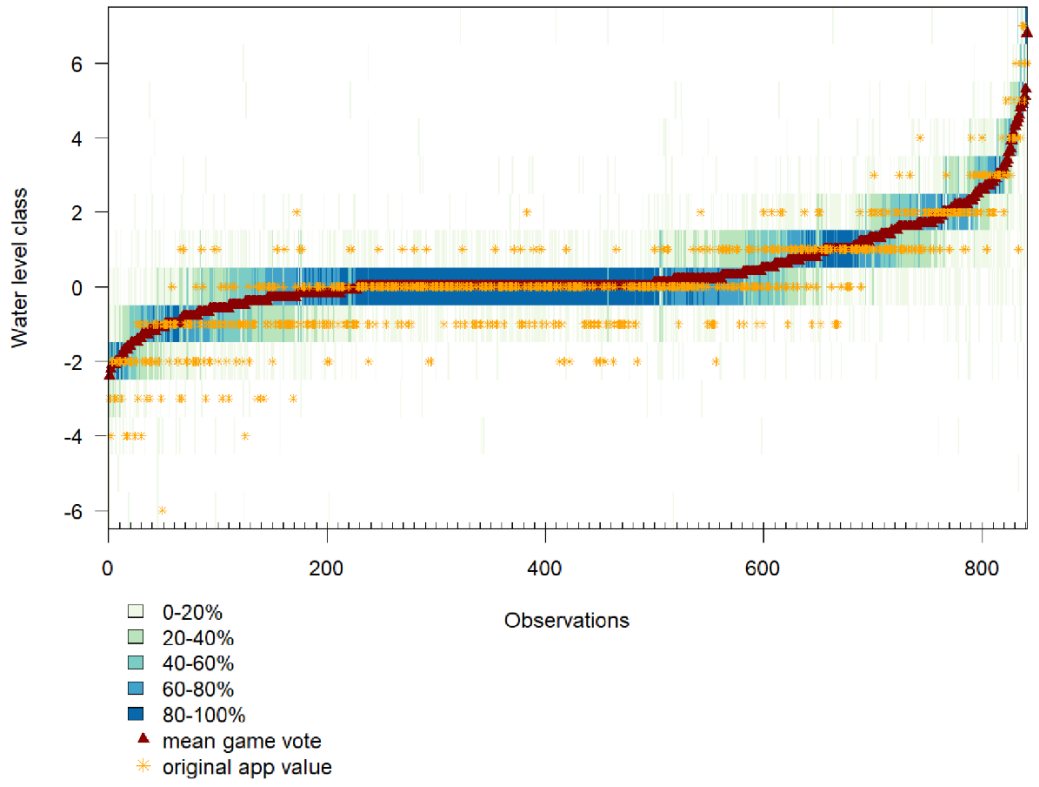


Figure 9: Agreement among players (in % of votes) per classified observation. Each column represents one observation, sorted according to the mean game vote (red triangle). Darker colours represent a higher agreement and lighter colours a lower agreement among the players. The original value of water level class submitted via the app is indicated by the orange star. Figure taken from Paper III.

In addition to correcting observations, the game can also increase the water level class resolution of the app submissions by taking the mean of all game votes. Therefore, the CrowdWater game provides further information for each observation, even for time series that have a well-placed staff gauge and good water level class estimates. For example, the Königseeache river has a nicely placed virtual staff gauge (see Figure 4) and only 19% of the observations were corrected by at least one class through the CrowdWater game and only one value was corrected by two classes (2% of all corrections). This is less than the 30% of observations that were corrected by at least one class across all locations in the CrowdWater game. However, 43% of the observations at Königseeache river have a value between two classes (*.3 ≤ mean game vote ≤ *.7, where * represents any class), which provides additional information compared to the app submission and contributes to a higher resolution (Figure 10).

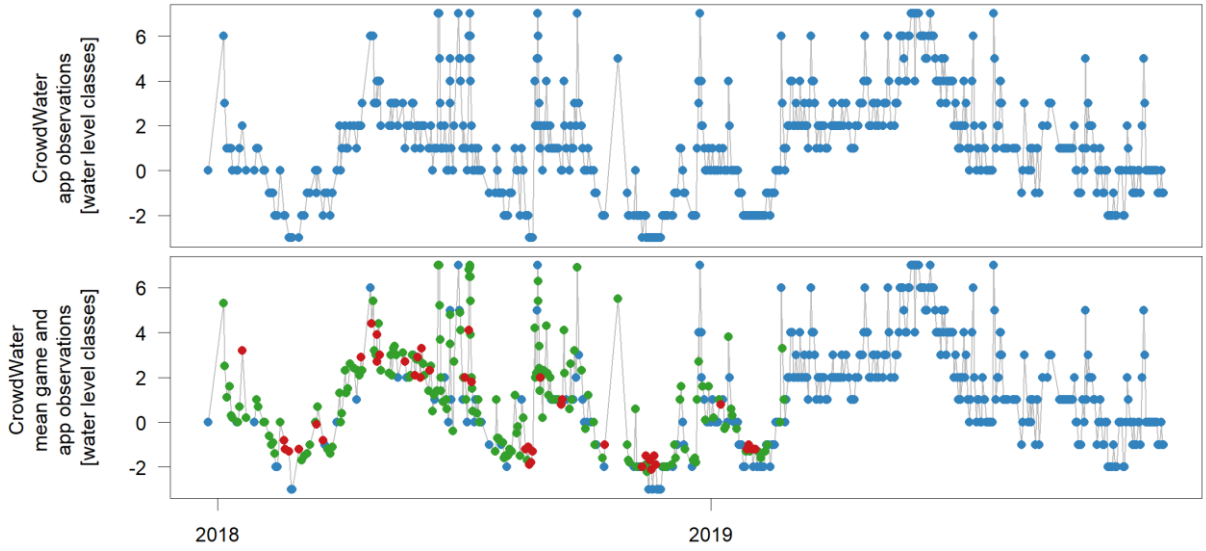


Figure 10: Crowdsourced water level class data for the Königseeache river. The upper plot shows the water level classes submitted via the app (blue) and the lower plot shows the mean water level classes from the game (green and red), which are supplemented by water level class observations from the app (blue) where the game has not yet provided 15 votes. The lower plot shows the increased water level class resolution that is possible by averaging CrowdWater game votes (green) as well as observations that were corrected by at least one class (red).

The results in Paper III show that regular players (> 24 classifications in the CrowdWater game, n = 58) had a significantly better median and a smaller range of the mean accuracy per player than novice players (≤ 24 classifications, n = 94). More CrowdWater game rounds improved the mean accuracy even further (Figure 11), showing that either the accuracy of players improves over time, or that players with a low accuracy are more likely to stop playing.

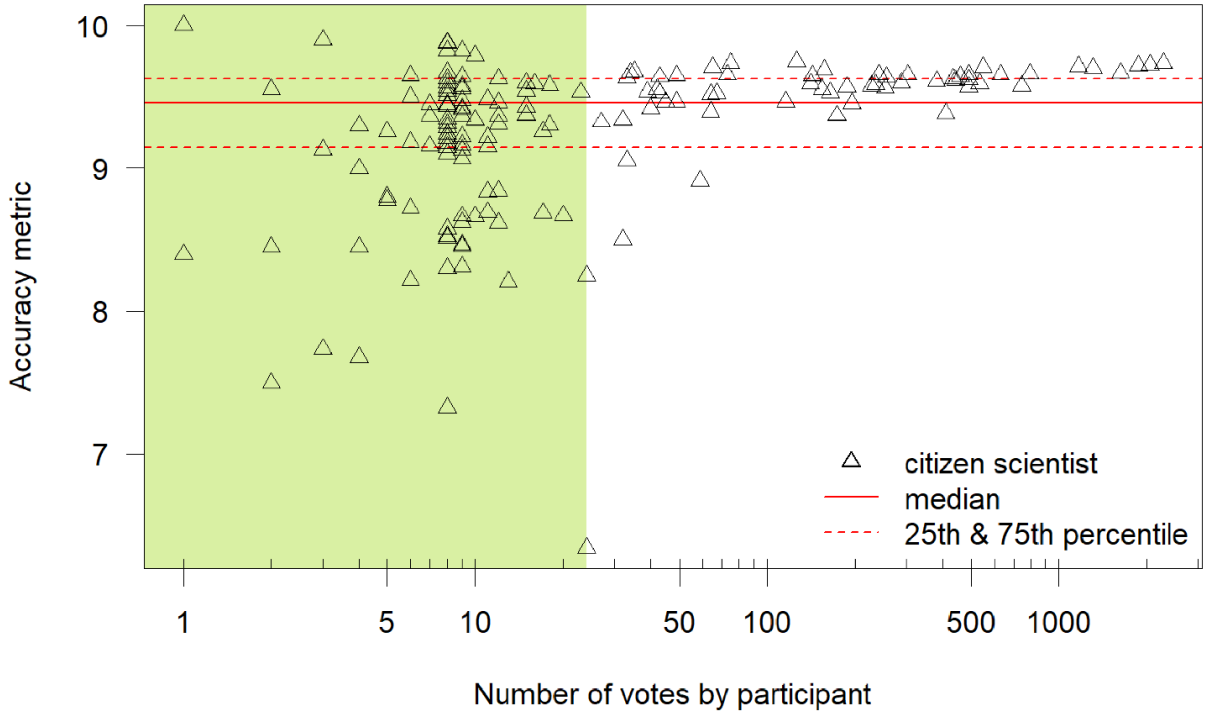


Figure 11: Mean accuracy per player as a function of the number of observations that that player classified. Each triangle represents one player. The lines indicate the median accuracy for all players (solid line) and the 25th and 75th percentile (dashed lines). The green shading indicates the novice players who played a maximum of two rounds (24 classifications). Note the log scale on the x-axis. Figure taken from Paper III.

5.4 Training citizen scientists

The results of the training study [*Paper IV*] suggest that many participants improved the placement of the virtual staff gauge after playing the CrowdWater game (i.e., after the training). The placement scores were significantly better after the training (Wilcoxon test; $p < 0.01$). Participants who had a low placement score before the training, i.e., participants who would potentially benefit more from a training, also performed significantly better after the training (Wilcoxon test; $p < 0.01$, Figure 12), although for 38% of participants the placement score was still low after the training.

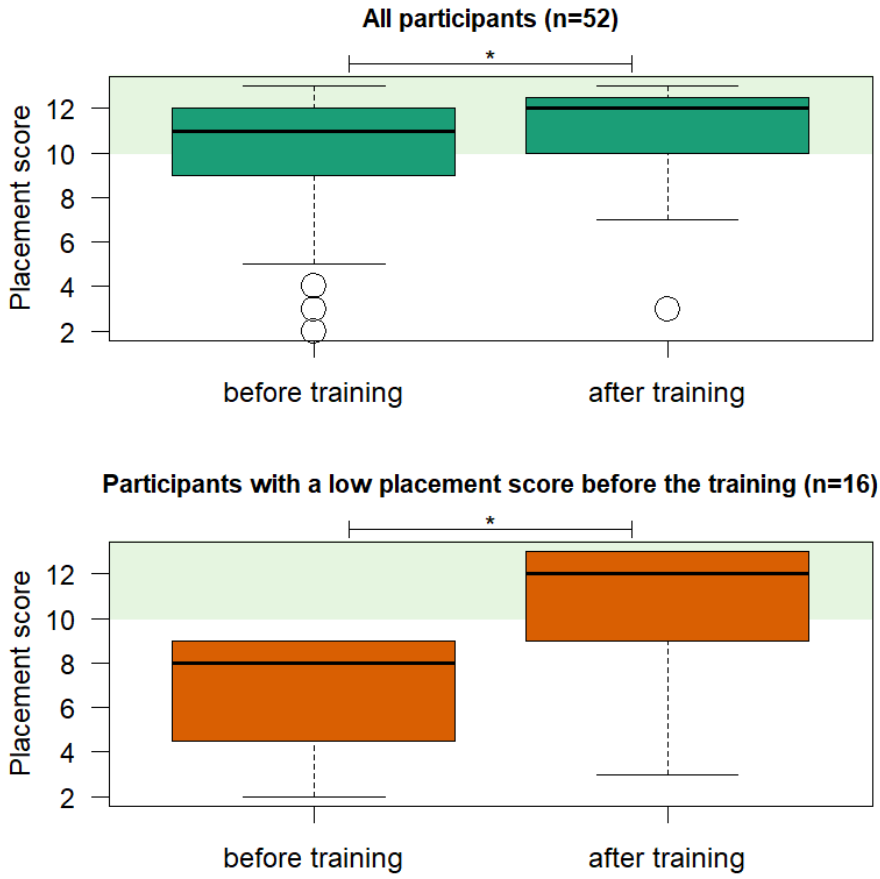


*Figure 12: Boxplots of the placement scores before the training and after the training for all participants (upper plot) and for participants who had a low placement score before the training (lower plot). There was a statistically significant difference in the placement scores before and after the training for both groups (indicated with the *) based on the Wilcoxon test ($p < 0.05$). The green shading indicates a good score (≥ 10 points). Figure adapted from Paper IV.*

The improvement was not necessarily related to the performance during the training. Participants with a good game score improved their placement score after the training, whereas participants with a low game score did not (Wilcoxon test; $p < 0.01$ and $p = 0.11$ respectively). Almost all participants with a good game score also had a good placement score after the training. However, also all participants with a low game score nonetheless had a good placement score after the training. Participants with a low placement score after the training mostly had an average game score. Due to the small number of participants in this category, this may not be a generalisable result (Figure 13).

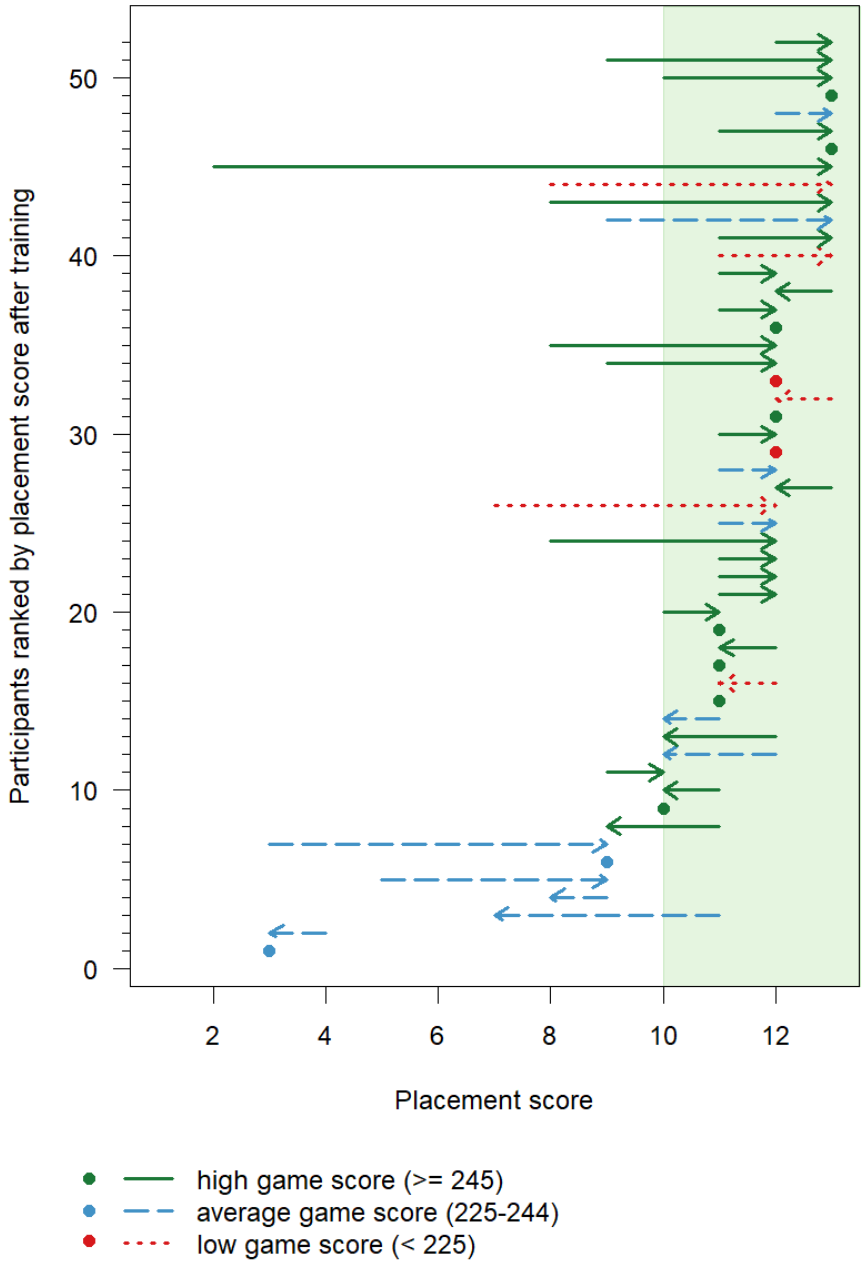


Figure 13: Placement scores before and after the training (x-axis); arrow points from before to after training score, dots indicate no change in the placement score. Each arrow or dot represents one participant (y-axis) and is coloured according to the game score they obtained during the training. Figure taken from Paper IV.

6 DISCUSSION

6.1 Is the virtual staff gauge concept suitable to crowdsource water level data?

Water level class data can be collected through a virtual staff gauge in a mobile smartphone application. Ideally, such observations are repeated at the same location, so that water level class time series are generated. These crowdsourced time series provide information about the river's dynamics (Figure 7).

Early tests showed that crowdsourcing streamflow data can be difficult and prone to errors [*Paper II*] and the value of such data for model calibration is limited [*Etter et al.*, 2018]. A more innovative approach was therefore needed, which resulted in the virtual staff gauge approach. Paper I shows that the concept of the virtual staff gauge is understandable and intuitive to most citizen scientists. Only about 10% of new staff gauge installations are problematic. Similarly, Paper IV shows that 36 of 52 participants (69%) placed the staff gauge well prior to receiving any training and that this increased to 44 of the 52 participants (85%) after training. Paper II shows that citizen scientists are better at estimating water level classes compared to streamflow. Only 13% of the passers-by chose a water level class that was more than one class off from the correct class and 48% chose the correct class. This was again confirmed in Paper III, which showed that for 70% of all app submissions, the app and mean game vote agreed and for only 10% of the observations the difference was more than one class. Aceves-Bueno et al. [2017] argue that keeping the skills of the citizen scientists in mind is crucial for a successful citizen science project. Therefore, we decided to use the virtual staff gauge approach, even though streamflow estimates correspond more to hydrological data that are traditionally used for model calibration or water resources allocation.

One of the main differences between CrowdWater and other citizen science projects that crowdsourced water level data, is that this project uses water level classes as opposed to a high-resolution staff gauge with absolute metric units [*Weeser et al.*, 2018; *Lowry et al.*, 2019]. We decided to use water level classes, because it is difficult to use an absolute scale in a virtual staff gauge. Having a virtual staff gauge that can be placed and retrieved on a smartphone application, enabled the approach to be fully scalable. Thus, staff gauges

could be placed by anybody anywhere without extra costs or needs for permits. A sticker-like virtual staff gauge that included a metric unit would, however, have been trickier to install, as on-site measurements would have been necessary [*Paper I*]. Therefore, we relied on virtual and relative water level class units.

Neither absolute high-resolution streamflow nor absolute high-resolution water level data are strictly necessary to calibrate a hydrological model, as classes provide enough information about the dynamics of streamflow responses [*Seibert and Vis*, 2016; *van Meerveld et al.*, 2017; *Etter et al.*, 2020]. Seibert and Vis [2016] calibrated 671 catchments with water level data, using the HBV model by optimizing the Spearman rank correlation coefficient [*Spearman*, 1904]. They showed that water level data can provide accurate model simulations, in particular in humid catchments. van Meerveld et al. [2017] presented a modelling study that used synthetic water level class data, i.e., streamflow that was converted into water level classes to calibrate the HBV model. The results showed that as few as two water level classes are already informative, however, five to seven classes improved the model performance even more. Etter et al. [2020] presented another modelling study with synthetic data. They also calibrated the HBV model with synthetic water level class data but included estimation errors (that correspond to the water level class estimation errors presented in Paper II) and assumed infrequent contribution times over the course of one year. The model still outperformed a lower benchmark with random parameter sets (i.e. without any water level or streamflow data input) and depending on the contribution times and error scenario was as good as the model calibrated with actual water level data. They show that 12 observations during one year already provide valuable information. This is considerably shorter compared to citizen science projects in other fields, such as ornithology, that require data over longer periods [*Sullivan et al.*, 2009; *Dickinson et al.*, 2010]. Etter et al. [2020] also show that although the modelling results are not very sensitive to errors, a lower estimation error is beneficial for the model calibration. Papers I and IV indicate that mistakes happen during the installation of the virtual staff gauge and Papers II and III show that errors also occur when citizen scientists estimate a water level class. Therefore, methods to improve the data quality of the water level class estimates are useful (see 6.2 Is it possible to crowdsource data quality control? and 6.3 Can a game for data quality control also be used to train new citizen scientists?).

In the future, it might be possible to use machine learning to determine the water level class in a photograph. However, at this stage we rather rely on citizen scientists to interpret the water level to avoid issues related to the precise location and angle at which the picture is taken [*Paper I*].

The limitations of the usefulness of water level class data for specific applications will have to be assessed in future studies. The geographic bias towards places where people live, might limit the amount of collected data in remote regions. The reduced temporal

resolution might be a problem for some data applications, as citizen scientists are unable to provide data at, for instance, hourly resolution. While the water level class observations are fairly accurate, some applications might require a higher accuracy and therefore, will not be able to use this virtual staff gauge based citizen science approach.

6.2 Is it possible to crowdsource data quality control?

Crowdsourcing data quality control, as opposed to expert quality control, is scalable and can easily be implemented in large-scale citizen science projects. During the early stages of the CrowdWater project, we could still check app submissions ourselves, however, the volume of app submissions quickly became too large, which led to the launch of the CrowdWater game. Similar approaches to crowdsource data quality control were implemented by iSpot [*Silvertown et al.*, 2015].

The data of the CrowdWater game show that it is possible to crowdsource data quality control. New observations for a water level time series (i.e., errors documented in Paper II) can be corrected or confirmed by showing picture pairs to multiple players and by taking their collective vote as the correct value. The agreement between players seems to be particularly high for the water level class zero (Figure 9). A picture of an observation with the water level class zero is very similar to the reference picture. Therefore, an estimate of the water level class might be easier for citizen scientists compared to extrapolating water level classes based on the reference features in the picture, such as stones on the streambank. Power et al. [2001] state that people intuitively look at similarities (in their case of maps) before they assess differences in patterns. Through the report function in the CrowdWater game, mistakes during the placement of the virtual staff gauge (i.e., errors documented in Paper I) can also be marked by the crowd. The results of Paper III are similar to the conclusions of other picture-based citizen science approaches, some of which were also gamified. Snapshot Serengeti, which asks participants to identify wildlife on photographs, reported 97.9% agreement between experts and the mean vote of participants, although the agreement varied according to species [*Swanson et al.*, 2016]. For Phylo, a game that tries to improve the alignment of the promoters of disease-related genes, it was shown that a citizen science approach can improve the accuracy of multiple genome sequence alignments [*Kawrykow et al.*, 2012]. Galaxy Zoo, a platform to visually classify galaxies, found that the classifications by citizen scientists were consistent with classifications by professional astronomers [*Lintott et al.*, 2008].

In addition to correcting erroneous data, the CrowdWater game also delivers a higher water level class resolution. This could be particularly important in cases, when the virtual staff gauge in the reference picture is too large, which is one of the most common mistakes (Table 2). A large staff gauge means that the water level remains within a few water level classes. Having too few classes might decrease the value of the data. Although

van Meerveld, et al. [2017] showed that as few as two water level classes were informative for model calibration, the performance of the model was better, when the water level data was split into five to seven classes. Therefore, for large virtual staff gauges the CrowdWater game can provide a better water level class resolution. However, even for staff gauges that appear properly sized, the game may result in higher resolution data, as shown for the Königseeache river, where 43% of the classified picture pairs suggested a half class (Figure 10).

6.3 Can a game for data quality control also be used to train new citizen scientists?

The idea of using the CrowdWater game also as a training tool developed through informal feedback from citizen scientists [*Paper IV*]. We also noticed improvements in the performance of players over time when analysing the CrowdWater game data [*Paper III*]. Figure 11 shows that players who participated for several rounds, generally had a higher voting accuracy than players who compared only a few picture pairs. Consequently, we analysed if the game could serve a second purpose.

The results of this study indicated that the CrowdWater game is a useful training tool for new citizen scientists. 63% of the participants who performed poorly prior to training placed the virtual staff gauge well after the training. The mistakes that were made when placing the staff gauge both before and after the training, resembled the mistakes presented in Paper I, which included making the staff gauge too big, not placing the zero line on the water level or placing the staff gauge with a distorting angle.

In total, 85% of all participants placed the staff gauge well after the training. This shows that the training did not help every participant. This might be attributed to the fact that the training with the CrowdWater game is an implicit approach, meaning that participants were not told what the relevant criteria for a good virtual staff gauge placement were. Most participants intuitively learned this by looking at many staff gauges and by trying to estimate water level classes both from well-placed and unfavourably-placed staff gauges. On the one hand, the benefit of such an approach is that crowdsourcing quality control and training can be mixed, whereas on the other hand some citizen scientists might have appreciated more explicit information and tips. Additionally, the CrowdWater game would likely be less fun to play, particularly for frequent players, if the same tips would be stated after every round. In order to better accommodate both types of citizen scientists, we recommend providing explicit tips on a project homepage. Newman et al. [2010] also recommend that different training approaches should be provided by citizen science projects.

Whether or not citizen scientists need to be trained varies vastly from project to project [*Gaddis*, 2018]. In some projects training can be crucial, such as CoCoRaHS, where citizen

scientists operate a meteorological station [*Reges et al.*, 2016] or a groundwater study in Canada, where citizen scientists measure water levels in wells [*Little et al.*, 2016]. However, when tasks can be achieved without any training, more citizen scientists might be inclined to join [*Paper IV*]. Keeping the barrier for entry low can, therefore, be beneficial. By offering a game as training, more people might go through this training, as the gamified interface can make it seem less like *"homework"*. So far, this training is not compulsory for citizen scientists before making their first observation with the CrowdWater app. Even though this could be adjusted in the future, we believe that participants are sufficiently self-motivated to play the CrowdWater game by the gamified features.

A combination of both approaches, training as well as crowdsourced data quality control, improves the data quality. Training ensures that fewer inadequate reference pictures are uploaded in the first place. This is important, as mistakes could always be missed by other citizen scientists during quality control. Additionally, fewer erroneous submissions would have to be removed, which saves time for the project administrators and avoids disappointing citizen scientists (when they see that their submission was incorrect or not very valuable).

Even with training, erroneous submissions are still uploaded, so that data quality control is still needed. For projects with a lot of submissions, crowdsourcing data quality control is often the only feasible approach, as expert review becomes too laborious [*Lintott et al.*, 2008; *Silvertown et al.*, 2015; *Freitag et al.*, 2016; *Kosmala et al.*, 2016; *Swanson et al.*, 2016].

7

CONCLUSIONS

This research has shown that water level class observations can be crowdsourced using a mobile smartphone application. The virtual staff gauge is an intuitive and scalable approach to collect data on water level dynamics. The virtual staff gauge can be placed anywhere in the world by citizen scientists. Thus, time series of water level classes can be collected wherever potential data users might require such data. Citizen scientists proved to be sufficiently motivated to contribute data on a regular basis and to collect valuable time series of hydrological data.

The water level class estimates by citizen scientists were more accurate than streamflow estimates, indicating that the virtual staff gauge is a suitable citizen science approach to collect water level data. Occasionally, citizen scientists still make mistakes, either when placing the virtual staff gauge or when estimating a new water level class for an existing location. These mistakes can be mitigated through quality control and training. Crowdsourcing data quality control based on the pictures of the rivers that are submitted with each observation through a game is effective at reducing errors and results in an increase of the water level class resolution. In addition, by playing this game, new citizen scientists are trained to place the virtual staff gauge. By observing different examples during the game, the players familiarised themselves with the virtual staff gauge approach and learned which sizes, angles and placements of virtual staff gauges were most suitable. This led to a better placement of the virtual staff gauge and thus better reference pictures for subsequent observations.

The resulting quality-controlled water level class data can potentially be used for hydrological model calibration, as shown in a preliminary study and other studies based on synthetic data (see 8. Outlook). Further studies should investigate under which conditions these data are most informative and should analyse the minimum requirements, in terms of time series length, number of observations and staff gauge placement for different catchment types and applications. The vote distribution collected through the game can help to estimate uncertainties associated with individual locations and measurements.

Hydrological measurements are scarce in many regions and citizen science is a valuable approach to supplement current hydrological data networks. Similar approaches can also

be used in many different fields and for various kinds of data. Some recommendations for future citizen science projects are:

- Consider carefully the types of observations that can be collected by citizen scientists. The difficulty and effort involved in the data collection dictates how many citizen scientists are likely to join and how much effort has to be invested in training.
- Evaluate what type of quality control can be implemented in the project and ensure that the collected data are of sufficiently high quality for the intended purpose. Photographs of observations often provide a simple and efficient basis to check the quality of the submitted observations.
- Plan sufficient time for community outreach and science communication, as citizen science projects need to be promoted and citizen scientists require regular feedback to ensure long-term participation.

8 OUTLOOK

8.1 Hydrological modelling with crowdsourced data

Although there are many ways to extend the analysis of the CrowdWater data, an obvious one is to determine its *"fitness-for-purpose"* and to use it for hydrological model calibration. Multiple studies have investigated the value of water level data for hydrological model calibration: Seibert and Vis [2016] calibrated a hydrological model (HBV) with water level measurements for over 600 catchments and found that especially in humid catchments the calibration with water level data worked surprisingly well. Mazzoleni et al. [2017] integrated synthetic crowdsourced water level data into a model and similarly found that such data can improve flood predictions.

Previous studies have also shown that water level class observations can in principle, be used to calibrate a hydrological model [*van Meerveld et al.*, 2017; *Etter et al.*, 2020]. These studies were, however, based on synthetic data and thus depended on assumptions, for example, the extent of the virtual staff gauge, the data quality and the frequency of the data collection. The errors of the synthetic observations were based on a previous field study with citizen scientists [*Paper II*]. In addition, the studies highlight that the value of the observations and the accuracy of the resulting model output depends on the accuracy of the observations.

8.1.1 Model calibration with real CrowdWater data

Potential future work can further extend these studies by 1) using the actually crowdsourced data for model calibration and 2) determining whether the quality-controlled game data lead to a better model performance than the data submitted directly via the app. Now that large amounts of crowdsourced water level class data have been collected, they can be used for future applied research. Therefore, it is important to develop methodologies to adequately assess the observation uncertainty. The described methodologies may also be useful for other citizen science projects.

We made a preliminary modelling study with the data from the Königseeache catchment (Figure 6, Figure 7 and Figure 10). At the time of the study, water level class data were available for the period December 2017 to August 2019. There were 486 water level class

observations, which means, on average, one water level class observation available per 1.2 days. Out of all these observations, 176 had received 15 or more votes in the CrowdWater game.

The catchment is 360 km^2 in size. Streamflow is measured by the Bavarian Environment Agency *"Berchtesgaden-Klärwerk"*, which is located roughly 10 km away from the CrowdWater location. The meteorological data were obtained from the German National Meteorological Service, as the majority of the catchment lies in Bavaria, Germany.

This preliminary study used the HBV model, as implemented in the HBV-light version [*Seibert and Vis*, 2012] to facilitate comparisons with previous studies [*Seibert and Vis*, 2016; *van Meerveld et al.*, 2017; *Etter et al.*, 2018, 2020]. The HBV model (Hydrologiska Byråns Vattenbalansavdelning) is a bucket-type runoff model [*Bergström*, 1976; *Lindström et al.*, 1997] that calculates snow, soil, groundwater and stream routing processes. The model can simulate streamflow when the input data (precipitation, temperature and evapotranspiration time series) are available.

The Spearman rank correlation coefficient was used as the objective function during the calibration with a genetic algorithm [*Seibert*, 2000]. Other commonly used objective functions were not possible since the water level class observations only reflect the dynamics of a stream, and do not provide volumetric information [*Spearman*, 1904; *Seibert and Vis*, 2016]. The calibrated model parameters were validated with streamflow measurements from the same time period. The resulting model efficiencies were compared to benchmark efficiencies, i.e., the maximum model performance obtained by calibrating the model with the streamflow data and the minimum performance obtained with random parameters. These benchmarks enable assessment of the results independent of the potential model structure uncertainty or uncertainties in the input data [*Seibert*, 2001; *Seibert et al.*, 2018].

The initial results show that the model calibrated with CrowdWater data (app and game) performed better than the lower benchmark. However, the results also suggest that the model fit is not perfect and that improvements are likely possible (Figure 14). Surprisingly, the models calibrated with the CrowdWater game and CrowdWater app data had a very similar performance. We believe that this is partly because the CrowdWater app data already have a high water level class resolution. It is possible that time series with fewer water level classes or with more erroneous data would benefit more from the additional information provided through the CrowdWater game. This should be investigated with further research.

The maximum uncertainty that is permissible in applied hydrology depends on the use of the model simulations. Even though higher uncertainties might be tolerated for some applications, it is crucial for data users to be aware of the quality of the data and the resulting model simulations. To better assess the usability of the data, and to assess if the CrowdWater data are *"fit-for-purpose"* [*Beven et al.*, 2012], potential applications of

hydrological data need to be analysed. Thus future data users will need to assess the fitness of the data for their specific purposes.

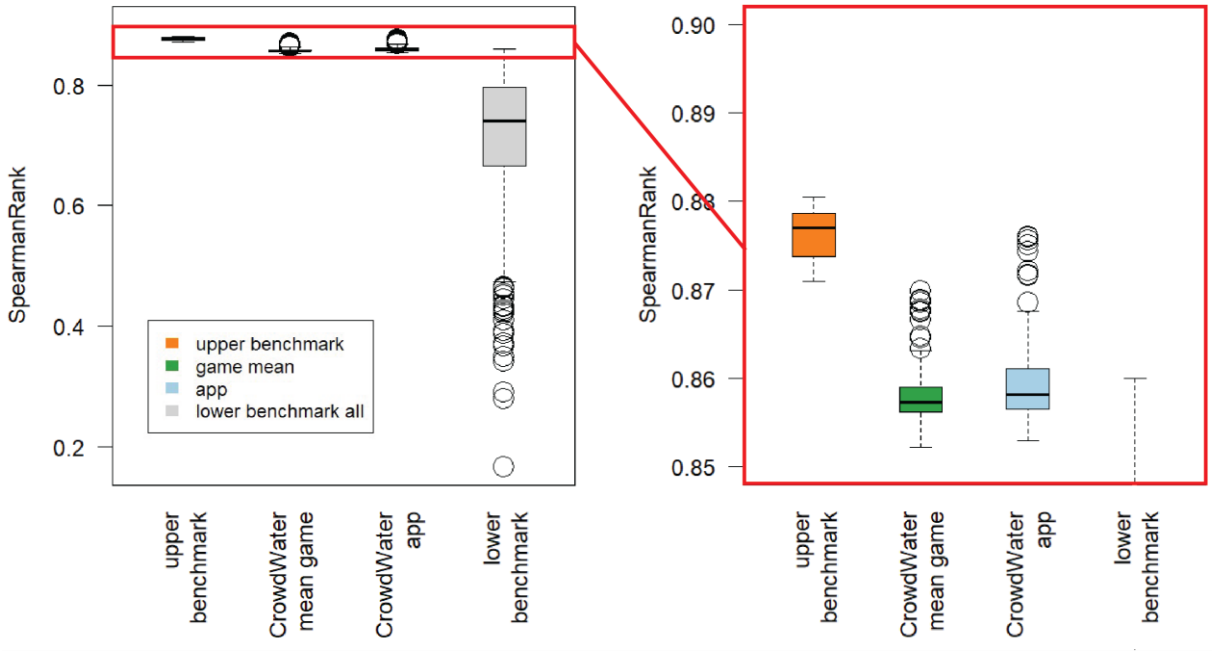


Figure 14: Boxplots of the values of the objective function for model validation for the same time period as the calibration. All model runs were calibrated by optimizing the Spearman Rank correlation coefficient. The model was calibrated 100 times. The left plot shows all resulting Spearman Rank values, whereas the right plot shows a zoomed in version, in order to emphasize the values of the upper benchmark, the CrowdWater game (mean game vote) and the CrowdWater app.

8.1.2 Including information on data uncertainty in model calibration

This preliminary study suggests that CrowdWater water level class data can be used in applied research where streamflow data are needed. Further improvements might still be possible, such as including the uncertainty information derived from the CrowdWater game, or including streamflow measurements or soil moisture observations.

The uncertainty of CrowdWater observations can be quantified based on the vote distribution of the CrowdWater game (i.e., the observation is more likely to be certain if the votes agree, and less certain if the votes disagree). The uncertainty estimate based on the spread of the votes for an observation in the CrowdWater game will enable data users (i.e., people who download the freely available CrowdWater data) to assess whether a specific observation (or a complete time series) has a relatively low or high overall uncertainty and to potentially exclude specific, highly uncertain observations. One of the earlier studies quantified the average accuracy of water level class estimates [*Paper II*]. This quantification likely overestimates the uncertainty as the survey only included first time users, whereas the observations in the app are generally submitted by frequent users.

Bootstrapping provides a means to assess how the uncertainty of the data affects the calibrated model parameters. Each observation gets 15-50 different votes in the game. The HBV model can be calibrated using a time series of randomly sampled (i.e.,

bootstrapped) votes from the game. By repeating this procedure multiple times and analysing the spread of the resulting streamflow simulations, the uncertainty in the model simulation due to the uncertainty in the input data can be quantified. Additionally, the bootstrapping results can be used to calculate a mean streamflow simulation (based on all bootstrapped simulations), thereby giving both an uncertainty range and a mean parameter set that is likely more robust against water level class estimation errors.

Future research could explore different strategies to include the available uncertainty information more directly into the modelling process, in order to assess how the overall model performance can be improved. This could be done by weighting the objective function during model calibration according to the uncertainty of each observation. The maximum likelihood framework, which is frequently used in statistics, provides a statistical approach for these weights. For instance, if the errors are assumed to be independent and Gaussian, each observation is weighted by the inverse of the observation error variance. However, because the observation error is not necessarily Gaussian, a different weighting might be more suitable, for example one that is based on higher order statistics, the range, or the interquartile range of the votes. Several of these metrics should be tested to investigate which metric works best for the CrowdWater game data. Such an approach has already been included in the HBV model to weight snow cover data between 0 and 1 according to the quality of the data (cloud cover on satellite image and overall quality of satellite image) [*unpublished, personal communication with Daphné Freudiger, 09.08.2019*]. We expect that accounting for the heterogeneity in the data quality significantly improves the resulting model quality and leads to a more realistic representation of the uncertainty in the model predictions.

These weighting metrics can only be calculated for observations that have already received a sufficient number of votes in the CrowdWater game. All other observations could instead be weighted by an average weight according for each water level class. The CrowdWater observations have different uncertainties depending on the water level. Generally, the uncertainties are highest for the high and low water level observations; observations close to water level class zero tend to be more certain (Figure 15).

The parameter uncertainty after calibration with CrowdWater water level classes can likely also be reduced through other means. One strategy is to use additional calibration data to ensure that the model interprets the CrowdWater observations adequately. Water level class data provide information regarding the dynamics of a hydrograph, but quantitative information on streamflow volumes is missing.

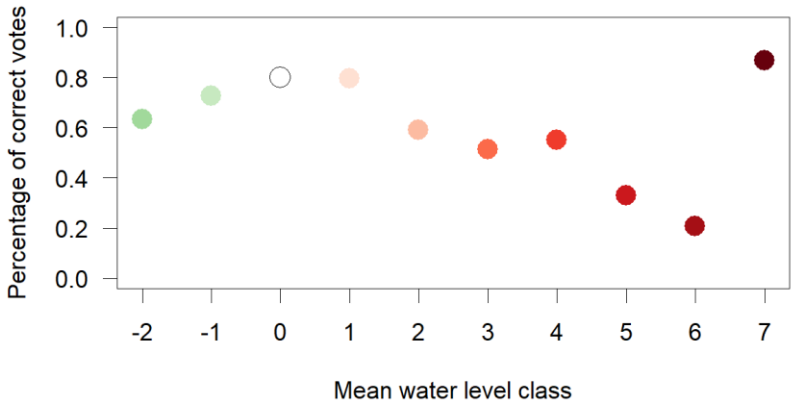


Figure 15: Percentage of correct votes per water level class (original app value) for all observations at the Königseeache river with the same mean water level class vote from the game. The colour of the dots indicates whether the observations were made at a time that the water level was lower (green), similar (white), or higher (red) than in the reference picture. The results for class 7 deviate from the overall pattern of lower agreement for more dissimilar water levels, because several observations had water levels significantly above class 7 and therefore gave the players certainty for which class to vote.

Seibert and Vis [2016] argued that particularly in humid catchments, the model partially derives the volumetric information from the precipitation input (as the potential and actual evapotranspiration are likely similar). In their study, they chose to couple water level observations with annual streamflow volumes to provide volumetric information. However, annual streamflow volumes are not collected through the CrowdWater project. Weeser et al. [2019] coupled water level observations with water balance information based on the precipitation and actual evapotranspiration estimates derived from remote sensing data. Another potential approach is to include a few high-accuracy streamflow measurements [*Buytaert et al.*, 2014; *Assumpção et al.*, 2018; *Pool and Seibert*, 2019]. An ongoing study investigating the combination of water level measurements with individual streamflow measurements has led to promising early results [*Pool and Seibert*, 2019]. This novel approach can be tested by merging water level class observations and a few streamflow measurements. The underlying assumption is that occasional streamflow measurements can be obtained by data users in parallel to CrowdWater observations with relatively little effort. This will not be necessary for every CrowdWater data user, but might be a useful approach for those who require smaller uncertainties and who can visit the location in question to make the measurements.

Another possibility is to use crowdsourced soil moisture observations (as collected by CrowdWater) as additional model input. If in a catchment with a specific water level class location also soil moisture observations are available, this information can be used to constrain the soil moisture parameters in the HBV model (or any other model). The overall model performance (as defined by the objective function value) might actually decrease due to the hydrograph not being exclusively fitted to water level class data, but the additional data might nonetheless help improve future predictions as the calibrated model parameters might be more consistent [*Seibert and McDonnell*, 2002]. A similar approach was used by Stahl et al. [2016], who calibrated the HBV model not only on streamflow measurements but also on snow and glacier cover data.

8.2 Other CrowdWater data

The CrowdWater app can be easily expanded and is already used to collect several other types of hydrological data (see 3.2 The CrowdWater app). Further research within this project aims to investigate the value of these data or to develop approaches for new observation types.

Flow information of intermittent streams can already be crowdsourced with the CrowdWater app and is currently a fairly popular observation type. Within the CrowdWater app, this category was named temporary streams, as this seemed more intuitive for citizen scientists. While many observations have been collected, so far, no analysis has been done on them. Further tests are needed to estimate the accuracy and consistency of these observations for a wide range of environments and the usefulness of these data to improve hydrological models. Some form of picture-based quality control might still be possible, at least for some sub-categories. For example, if some water is clearly visible in a photograph, the category *"dry riverbed"* is unlikely to be correct, although it might be difficult to distinguish between trickling and flowing water in a photograph. Further research is needed to assess, if such a limited form of data quality control is helpful.

The CrowdWater app currently does not collect any data on water quality, however, this could potentially be included. In order to assess the usefulness of such data for water quality modelling, research with synthetic data will first need to be conducted, similar to Seibert and Beven [2009], Seibert and Vis [2016], van Meerveld et al. [2017], Etter et al. [2018]. The research with the synthetic data can focus on questions, such as: what is the best balance between spatial and temporal resolution of crowdsourced water quality data, what is the best balance between accuracy and amount of data and what is the potential value of inexpensive sensors for citizen science approaches. Even for this observation type some limited form of the CrowdWater game might be possible, as some variables of water quality, such as algal blooms or oil spills are visible. Further research could investigate the feasibility of visual data quality control for water quality observations.

8.3 CrowdWater game

Currently, when there is a discrepancy between the original app value and the mean game vote, the mean game vote is better in 74% of cases. In addition to the mean vote, the CrowdWater game also provides a vote distribution per observation. If a higher accuracy is needed, this vote distribution (e.g., a bimodal distribution or a large spread of votes) could be used to gain further information about this observation, which might be used to distinguish observations where the mean game vote was incorrect after all.

Alternatively, trusted citizen scientists (i.e., citizen scientists with a high accuracy score in the CrowdWater game) could be asked to assess discrepancies between the CrowdWater game and the app or picture pairs where the vote distribution indicates a large uncertainty.

The CrowdWater game could be expanded to provide more explicit training material for new citizen scientists. A separate interface could be developed for new and regular citizen scientists, so that the quality control is not interrupted for regular players. In a separate interface, new citizen scientists could receive tips on how to place the virtual staff gauge well and what reference structures to look for. This should, in particular, be included if the CrowdWater game is made into a mandatory training for new citizen scientists.

ACKNOWLEDGEMENTS

I would like to express my gratitude towards everybody that made this thesis and the last four years of my Ph.D. studies possible!

Jan Seibert and Ilja van Meerveld initiated the CrowdWater project and hired me as their Ph.D. student - thereby starting this wonderful journey. Both have taught me how to approach scientific studies, how to write publications and how to present at conferences. They were always very patient and encouraging with their feedback. I would also like to thank Jan for always having an open mind towards new CrowdWater ideas, which really made me feel like my input was appreciated. I would like to thank Ilja for our weekly meetings and for often nudging me in the right directions, thereby making the publication process much smoother. Thank you to both Jan and Ilja for taking a lot of time to develop and facilitate such a welcoming, friendly and outgoing group.

Ross Purves, my third committee member, always provided excellent feedback and guidance during the committee meetings. He was always looking out for me as a student and made sure that I could complete my Ph.D. in time.

Simon Etter was a great project team mate. I am very happy that the two of us were able to shape the CrowdWater project together. I really appreciate the open and uncomplicated work relationship that we developed. He gave me the opportunity to freely share ideas with him and through this open communication, I believe the project benefited a lot. I really appreciate that he was always ready to help out whenever I needed something. I had a lot of fun during our field trips: whether that was a stroll through Sihlwald or surveying Brugg's residents in freezing temperatures. I also really enjoyed developing the CrowdWater logo and Droppy graphics together, our joint conferences, making CrowdWater videos together and the numerous other joint adventures. It was a pleasure to share our Ph.D. experience!

The H2K group was an excellent research group and I thoroughly enjoyed spending my roughly 10.000 coffee breaks and 5.000 lunches with them. The atmosphere was always light-hearted and friendly, which was great to switch off during the communal breaks. I also appreciate that I could always ask anyone for help, whether scientific or private. Many contributed to the CrowdWater project and provided invaluable feedback in the

early stages of the project! I would also like to thank them for the feedback on all my presentations and posters and the lively discussions on how to improve them.

I had wonderful office mates during the last four years: Tobias Bolch, Michal Jeníček, Ling Wang, Anna Sikorska-Senoner, Rosy Lane and Carolina Natel de Moura. Everyone maintained a welcoming office atmosphere with the occasional chat, lots of tea drinking and a visible productivity that motivated me as well.

Marc Vis tirelessly explained HBV to me, fixed my R code, answered my computer questions and of course organised the H2K Christmas party and SOLA. I really appreciate how much he encouraged us to ask questions and thereby undoubtedly saved me from many frustrating hours.

Ross Purves and Isabelle Gärtner-Roer organised the Graduate School. During the courses and retreats we did not only learn how to do a Ph.D. but found many friends all over GIUZ, who were on the same journey as us.

Our incredible citizen scientists have contributed over 10.000 observations to the CrowdWater project, provided valuable feedback and made the project work very enjoyable. We were thrilled that we were able to reach so many users and are delighted that many seem to genuinely enjoy contributing on a regular basis. This thesis would not have been possible without their many contributions!

Philipp Hummer and his colleagues at SPOTTERON (www.spotteron.net) did a great job in developing the CrowdWater app and the CrowdWater game. They always took great care to design everything in a user friendly way, for which I am sure all CrowdWater citizen scientists are thankful as well.

Thank you to the Swiss National Science Foundation (SNSF), which funded this study (www.snf.ch; project 163008, CrowdWater).

Finally, I would like to thank my family. My parents instilled a passion for Geography in me from a very early age and were very supportive of all my decisions. Throughout the last four years they were always there with a helping hand. I also really appreciate the support by Hemma, Wolfgang and Wolfi. I would also like to thank Michael for his love, support, encouragement and for our joint activities during the last few years.

REFERENCES

Aceves-Bueno, E., A. S. Adeleye, M. Feraud, Y. Huang, M. Tao, Y. Yang, and S. E. Anderson (2017), The Accuracy of Citizen Science Data: A Quantitative Review, *Bull. Ecol. Soc. Am.*, *98*(4), 278–290, doi:10.1002/bes2.1336.

Allen, D. C., D. A. Kopp, K. H. Costigan, T. Datry, B. Hugueny, D. S. Turner, G. S. Bodner, and T. J. Flood (2019), Citizen scientists document long-term streamflow declines in intermittent rivers of the desert southwest, USA, *Freshw. Sci.*, *38*(2), 244–256, doi:10.1086/701483.

Aono, Y., and K. Kazui (2008), Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century, *Int. J. Climatol.*, *28*(7), 905–914, doi:10.1002/joc.1594.

Assumpção, T. H., I. Popescu, A. Jonoski, and D. P. Solomatine (2018), Citizen observations contributing to flood modelling: opportunities and challenges, *Hydrol. Earth Syst. Sci.*, *22*, 1473–1489, doi:10.5194/hess-2017-456.

August, T. A., S. E. West, H. Robson, J. Lyon, J. Huddart, L. F. Velasquez, and I. Thornhill (2019), Citizen meets social science: Predicting volunteer involvement in a global freshwater monitoring experiment, *Freshw. Sci.*, *38*(2), 321–331, doi:10.1086/703416.

Barras, H., A. Hering, A. Martynov, P.-A. Noti, U. Germann, and O. Martius (2019), Experiences with >50'000 crowd-sourced hail reports in Switzerland, *Am. Meteorol. Soc.*, 1–35, doi:10.1175/bams-d-18-0090.1.

Bellinger, D. C. (2016), Lead Contamination in Flint - An Abject Failure to Protect Public Health, *N. Engl. J. Med.*, *363*(1), 1–3, doi:10.1056/NEJMp1002530.

Bergeron, T. (1949), The problem of artificial control of rainfall on the globe. Part II: The coastal orographic maxima of precipitation in autumn and winter, *Tellus*, *1*, 15–32, doi:10.1111/j.2153-3490.1949.tb01264.x.

Bergeron, T. (1960), Operation and results of '"Project Pluvius.,"' in *Physics of Precipitation, Geophys. Monogr., No. 5*, pp. 152–157, Amer. Geophys. Union.

Bergström, S. (1976), *Development and application of a conceptual runoff model for Scandinavian catchments*, edited by S. Bergström, SMHI Norrköping, Report RH07, Norrköping, Sweden.

Beven, K., W. Buytaert, and L. A. Smith (2012), On virtual observatories and modelled realities (or why discharge must be treated as a virtual variable), *Hydrol. Process.*, *26*(12), 1906–1909, doi:10.1002/hyp.9261.

Beven, K. J. (2012), *Rainfall-Runoff Modelling: The Primer*, 2nd ed., Wiley-Blackwell,

Chichester UK.

Bishop, K., I. Buffam, M. Erlandsson, J. Fölster, H. Laudon, J. Seibert, and J. Temnerud (2008), Aqua Incognita: the unknown headwaters, *Hydrol. Process.*, *22*, 1239–1242, doi:10.1002/hyp.7049.

Bonney, R., C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk (2009a), Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy, *Bioscience*, *59*(11), 977–984, doi:10.1525/bio.2009.59.11.9.

Bonney, R., H. Ballard, R. Jordan, E. McCallie, T. Phillips, J. Shirk, and C. C. Wilderman (2009b), *Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education*, Washington, D.C.

Bonney, R., J. L. Shirk, T. B. Phillips, A. Wiggins, H. L. Ballard, A. J. Miller-Rushing, and J. K. Parrish (2014), Next Steps for Citizen Science, *Science*, *343*(6178), 1436–1437, doi:10.1126/science.1251554.

Bonney, R., C. B. Cooper, and H. Ballard (2016), The theory and practice of citizen science: Launching a new journal, *Citiz. Sci. Theory Pract.*, *1*(1), 1–1, doi:dx.doi.org/10.5334/cstp.65.

Bowser, A. E., and A. Wiggins (2015), Privacy in Participatory Research : Advancing Policy to support Human Computation, *Hum. Comput.*, *2*(1), 19–44, doi:10.15346/hc.v2i1.3.

Breuer, L., N. Hiery, P. Kraft, M. Bach, A. H. Aubert, and H.-G. Frede (2015), HydroCrowd: a citizen science snapshot to assess the spatial control of nitrogen solutes in surface waters, *Sci. Rep.*, *5*(16503), doi:10.1038/srep16503.

Burgess, H. K., L. B. DeBey, H. E. Froehlich, N. Schmidt, E. J. Theobald, A. K. Ettinger, J. HilleRisLambers, J. Tewksbury, and J. K. Parrish (2016), The science of citizen science: Exploring barriers to use as a primary research tool, *Biol. Conserv.*, doi:10.1016/j.biocon.2016.05.014.

Buytaert, W. et al. (2014), Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development, *Front. Earth Sci.*, *2*(26), 21, doi:10.3389/feart.2014.00026.

Castilla, E. P., D. G. F. Cunha, F. W. F. Lee, S. Loiselle, K. C. Ho, and C. Hall (2015), Quantification of phytoplankton bloom dynamics by citizen scientists in urban and peri-urban environments, *Environ. Monit. Assess.*, *187*(690), doi:10.1007/s10661-015-4912-9.

Catlin-Groves, C. L. (2012), The Citizen Science Landscape: From Volunteers to Citizen Sensors and Beyond, *Int. J. Zool.*, *2012*, 1–14, doi:10.1155/2012/349630.

Chao, L., Z. Hui, and Z. Xiaofeng (2015), Data quality assessment in hydrological information systems, *J. Hydroinformatics*, 640–661, doi:10.2166/hydro.2015.042.

Clery, D. (2011), Galaxy Zoo Volunteers Share Pain and Glory of Research, *Science*, *333*(6039), 173–175.

Cooper, C. (2016), *Citizen Science: How Ordinary People Are Changing the Face of Discovery*, The Overlook Press, Peter Mayer Publishers, Inc., New York.

Cooper, S., F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. Players (2010), Predicting protein structures with a multiplayer online game, *Nature*, *466*(7307), 756–760, doi:10.1038/nature09304.

Le Coz, J. et al. (2016), Crowdsourced data for flood hydrology : feedback from recent citizen science projects in Argentina , France and New Zealand, *J. Hydrol.*, *541*, 766–777, doi:10.1016/j.jhydrol.2016.07.036.

Crall, A. W., R. Jordan, K. Holfelder, G. J. Newman, J. Graham, and D. M. Waller (2013), The impacts of an invasive species citizen science training program on participant attitudes, behavior, and science literacy, *Public Underst. Sci.*, *22*(6), 745–764, doi:10.1177/0963662511434894.

Crowston, K., and N. R. Prestopnik (2013), Motivation and data quality in a citizen science game: A design science evaluation, *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 450–459, doi:10.1109/HICSS.2013.413.

Danielsen, F. et al. (2014), A Multicountry Assessment of Tropical Resource Monitoring by Local Communities, *Bioscience*, *64*(3), 236–251, doi:10.1093/biosci/biu001.

Davids, J. C., N. Devkota, A. Pandey, R. Prajapati, B. A. Ertis, M. M. Rutten, S. W. Lyon, T. A. Bogaard, and N. van de Giesen (2019), Soda Bottle Science—Citizen Science Monsoon Precipitation Monitoring in Nepal, *Front. Earth Sci.*, *7*(46), doi:10.3389/feart.2019.00046.

Dickerson-Lange, S. E., K. B. Eitel, L. Dorsey, T. E. Link, and J. D. Lundquist (2016), Challenges and successes in engaging citizen scientists to observe snow cover: From public engagement to an educational collaboration, *J. Sci. Commun.*, *15*(1), 1–14, doi:10.22323/2.15010201.

Dickinson, J. L., B. Zuckerberg, and D. N. Bonter (2010), Citizen Science as an Ecological Research Tool: Challenges and Benefits, *Annu. Rev. Ecol. Evol. Syst.*, *41*(1), 149–172, doi:10.1146/annurev-ecolsys-102209-144636.

Dickinson, J. L., J. Shirk, D. Bonter, R. Bonney, R. L. Crain, J. Martin, T. Phillips, and K. Purcell (2012), The current state of citizen science as a tool for ecological research and public engagement, *Front. Ecol. Environ.*, *10*(6), 291–297, doi:10.1890/110236.

Dunn, E. H., C. M. Francis, P. J. Blancher, S. R. Drennan, M. A. Howe, D. Lepage, C. S. Robbins, K. V. Rosenberg, J. R. Sauer, and K. G. Smith (2005), Enhancing the scientific value of the christmas bird count, *Auk*, *122*(1), 338–346, doi:10.1642/0004-8038(2005)122[0338:ETSVOT]2.0.CO;2.

Eitzel, M. V et al. (2017), Citizen Science Terminology Matters: Exploring Key Terms, *Citiz. Sci. Theory Pract.*, *2*(1), 1, doi:10.5334/cstp.96.

Emmerik, T. Van, and A. Schwarz (2020), Plastic debris in rivers, *WIREs Water*, *7*(e1398), doi:10.1002/wat2.1398.

Engel, S. R., and J. R. Voshell (2002), Volunteer biological monitoring: can it accurately assess the ecological condition of streams?, *Am. Entomol.*, *48*, 164–177, doi:citeulike-article-id:9349617.

Etter, S., B. Strobl, J. Seibert, and I. van Meerveld (2018), Value of uncertain streamflow observations for hydrological modelling, *Hydrol. Earth Syst. Sci.*, *22*, 5243–5257, doi:10.5194/hess-22-5243-2018.

Etter, S., B. Strobl, J. Seibert, and H. J. Meerveld (2020), Value of crowd-based water level class observations for hydrological model calibration, *Water Resour. Res.*, doi:10.1029/2019WR026108.

Etter, S., B. Strobl, J. Seibert, H. J. van Meerveld, and K. Niebert (in review), What motivates

people to participate in environmental citizen science projects?, *Citiz. Sci. Theory Pract.*

Fekete, B. M., U. Looser, A. Pietroniro, and R. D. Robarts (2012), Rationale for Monitoring Discharge on the Ground, *J. Hydrometeorol.*, *13*(6), 1977–1986, doi:10.1175/JHM-D-11-0126.1.

Fienen, M. N., and C. S. Lowry (2012), Social.Water—A crowdsourcing tool for environmental data acquisition, *Comput. Geosci.*, *49*, 164–169, doi:10.1016/j.cageo.2012.06.015.

Freitag, A., R. Meyer, and L. Whiteman (2016), Strategies Employed by Citizen Science Programs to Increase the Credibility of Their Data, *Citiz. Sci. Theory Pract.*, *1*(1), 1–11, doi:10.5334/cstp.6.

Gaddis, M. (2018), Training Citizen Scientists for Data Reliability: a Multiple Case Study to Identify Themes in Current Training Initiatives, University of the Rockies.

Ganzevoort, W., R. J. G. van den Born, W. Halffman, and S. Turnhout (2017), Sharing biodiversity data: citizen scientists? concerns and motivations, *Biodivers. Conserv.*, doi:10.1007/s10531-017-1391-z.

Goodchild, M. F. (2007), Citizens as sensors: The world of volunteered geography, *GeoJournal*, *69*(4), 211–221, doi:10.1007/s10708-007-9111-y.

Graham, E. A., S. Henderson, and A. Schloss (2011), Using Mobile Phones to Engage Citizen Scientists in Research, *Eos (Washington. DC).*, *92*(38), 313–315, doi:10.1029/2011EO380002.

Haklay, M. (2010), How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets, *Environ. Plan. B Plan. Des.*, *37*(4), 682–703, doi:10.1068/b35097.

Haklay, M. (2013), Citizen Science and Volunteered Geographic Information - overview and typoology of participation, in *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, edited by D. Sui, S. Elwood, and M. Goodchild, pp. 105–122, Springer, Berlin.

Haklay, M. (2014), Citizen Science in Oxford English Dictionary, Available from: https://povesham.wordpress.com/2014/09/10/citizen-science-in-oxford-english-dictionary/?blogsub=confirming#blog_subscription-3 (Accessed 15 December 2019)

Haklay, M. (2015), *Citizen Science and Policy : A European Perspective*.

Haklay, M. M., S. Mazumdar, and J. Wardlaw (2018), Citizen Science for Observing and Understanding the Earth, in *Earth Observation Open Science and Innovation*, edited by P. Mathieu and C. Aubrecht, pp. 69–88, Springer, Cham.

Hendriks, M. R. (2010), *Introduction to Physical Hydrology*, Oxford University Press Inc., New York.

Hennon, C. C. et al. (2015), Cyclone center: can citizen scientists improve tropical cyclone intensity records?, *Am. Meteorol. Soc.*, *96*(4), 591–608, doi:10.1175/BAMS-D-13-00152.1.

Hill, D. F., G. J. Wolken, K. W. Jones, R. Crumley, and A. Arendt (2018), Crowdsourcing snow depth data with citizen scientists, *Eos (Washington. DC).*, *99*, doi:10.1029/2018EO108991.

Hornberger, G. M., J. P. Raffensperger, P. L. Wiberg, and K. N. Eshleman (1998), *Elements of Physical Hydrology*, The John Hopkins University Press, Baltimore, Maryland.

Irwin, A. (1995), *Citizen Science - A Study of People, Expertise and Sustainable Development*, Routledge, New York.

Jennett, C. et al. (2016), Motivations, learning and creativity in online citizen science, *J. Sci. Commun.*, *15*(3), 1–23.

Jollymore, A., M. J. Haines, T. Satterfield, M. S. Johnson, T. Satter, and M. S. Johnson (2017), Citizen science for water quality monitoring : Data implications of citizen perspectives, *J. Environ. Manage.*, *200*, 456–467, doi:10.1016/j.jenvman.2017.05.083.

Jones, T. et al. (2018), Massive Mortality of a Planktivorous Seabird in Response to a Marine Heatwave, *Geophys. Res. Lett.*, *45*, 3193–3202, doi:10.1002/2017GL076164.

Kampf, S., B. Strobl, J. Hammond, A. Annenberg, S. Etter, C. Martin, K. Puntenney-Desmond, J. Seibert, and I. van Meerveld (2018), Testing the waters: Mobile apps for crowdsourced streamflow data, *Eos (Washington. DC).*, *99*, doi:10.1029/2018EO096355.

Kawrykow, A., G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, L. Sarmenta, M. Blanchette, and J. Waldispühl (2012), Phylo: A citizen science approach for improving multiple sequence alignment, *PLoS One*, *7*(3), doi:10.1371/journal.pone.0031362.

Kiang, J. E. et al. (2018), A Comparison of Methods for Streamflow Uncertainty Estimation, *Water Resour. Res.*, doi:10.1029/2018WR022708.

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, *42*, W03S04, doi:10.1029/2005WR004362.

Kloetzer, L., D. Schneider, and J. Costa (2016), Not So Passive: Engagement and Learning in Volunteer Computing, *Hum. Comutation J.*, 25–68, doi:10.15346/hc.v3i1.4.

Koch, J., and S. Stisen (2017), Citizen science: A new perspective to advance spatial pattern evaluation in hydrology, *PLoS One*, *12*(5), 1–20, doi:10.1371/journal.pone.0178165.

Kosmala, M., A. Wiggins, A. Swanson, and B. Simmons (2016), Assessing data quality in citizen science, *Front. Ecol. Environ.*, *14*(10), 551–560, doi:10.1002/fee.1436.

Kundzewicz, Z. W. (1997), Water resources for sustainable development, *Hydrol. Sci. J.*, *42*(4), 467–480, doi:10.1080/02626669709492047.

Lindström, G., B. Johansson, M. Persson, M. Gardelin, and S. Bergström (1997), Development and test of the distributed HBV-96 hydrological model, *J. Hydrol.*, *201*(1–4), 272–288, doi:10.1016/S0022-1694(97)00041-3.

Lintott, C. J. et al. (2008), Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey, *Mon. Not. R. Astron. Soc.*, *389*(3), 1179–1189, doi:10.1111/j.1365-2966.2008.13689.x.

Little, K. E., M. Hayashi, and S. Liang (2016), Community-Based Groundwater Monitoring Network Using a Citizen-Science Approach, *Groundwater*, *54*(3), 317–324, doi:10.1111/gwat.12336.

Loiselle, S. A., D. Gasparini, F. Cunha, S. Shupe, E. Valiente, L. Rocha, E. Heasley, P. Pérez Belmont, and A. Baruch (2016), Micro and Macroscale Drivers of Nutrient

Concentrations in Urban Streams in South, Central and North America, *PLoS One*, *11*(9), 1–16, doi:10.1371/journal.pone.0162684.

Lottig, N. R., T. Wagner, E. N. Henry, K. S. Cheruvelil, K. E. Webster, J. A. Downing, and C. A. Stow (2014), Long-Term Citizen-Collected Data Reveal Geographical Patterns and Temporal Trends in Lake Water Clarity, *PLoS One*, *9*(4), doi:10.1371/journal.pone.0095769.

Lowry, C. S., and M. N. Fienen (2013), CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists, *Ground Water*, *51*(1), 151–156, doi:10.1111/j.1745-6584.2012.00956.x.

Lowry, C. S., M. N. Fienen, D. M. Hall, and K. F. Stepenuck (2019), Growing Pains of Crowdsourced Stream Stage Monitoring Using Mobile Phones: The Development of CrowdHydrology, *Front. Earth Sci.*, *7*(128), 1–10, doi:10.3389/feart.2019.00128.

Manning, R. (1891), On the flow of water in open channels and pipes, *Trans. Inst. Civ. Eng. Irel.*, *20*, 161–207.

Matthias, S., M. Vitos, J. Altenbuchner, G. Conquest, J. Lewis, and M. Haklay (2014), Taking Participatory Citizen Science to Extremes, *IEEE Pervasive Comput.*, *13*(2), 20–29, doi:10.1109/MPRV.2014.37.

Mazzoleni, M., M. Verlaan, L. Alfonso, M. Monego, D. Norbiato, M. Ferri, and D. P. Solomatine (2017), Can assimilation of crowdsourced data in hydrological modelling improve flood prediction?, *Hydrol. Earth Syst. Sci.*, *21*(2), 839–861, doi:10.5194/hess-21-839-2017.

Mazzoleni, M., V. Juliette Cortes Arevalo, U. Wehn, L. Alfonso, D. Norbiato, M. Monego, M. Ferri, and D. P. Solomatine (2018), Exploring the influence of citizen involvement on the assimilation of crowdsourced observations: A modelling study based on the 2013 flood event in the Bacchiglione catchment (Italy), *Hydrol. Earth Syst. Sci.*, *22*, 391–416, doi:10.5194/hess-22-391-2018.

McKinley, D. C. et al. (2017), Citizen science can improve conservation science, natural resource management, and environmental protection, *Biol. Conserv.*, *208*, 15–28, doi:10.1016/j.biocon.2016.05.015.

McMillan, H., T. Krueger, and J. Freer (2012), Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, *Hydrol. Process.*, *26*(26), 4078–4111, doi:10.1002/hyp.9384.

van Meerveld, H. J., M. J. P. Vis, and J. Seibert (2017), Information content of stream level class data for hydrological model calibration, *Hydrol. Earth Syst. Sci.*, *21*(9), 4895–4905, doi:10.5194/hess-21-4895-2017.

Mulligan, M. (2013), WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally, *Hydrol. Res.*, *44*(5), 748, doi:10.2166/nh.2012.217.

Newman, G., A. Crall, M. Laituri, J. Graham, T. Stohlgren, J. C. Moore, K. Kodrich, and K. A. Holfelder (2010), Teaching citizen science skills online: Implications for invasive species training programs, *Appl. Environ. Educ. Commun.*, *9*(4), 276–286, doi:10.1080/1533015X.2010.530896.

Newman, G., A. Wiggins, A. Crall, E. Graham, S. Newman, and K. Crowston (2012), The future of Citizen science: Emerging technologies and shifting paradigms, *Front. Ecol. Environ.*, *10*(6), 298–304, doi:10.1890/110294.

Njue, N., J. Stenfert Kroese, J. Gräf, S. R. Jacobs, B. Weeser, L. Breuer, and M. C. Rufino (2019), Citizen science in hydrological monitoring and ecosystem services management: State of the art and future prospects, *Sci. Total Environ.*, *693*, 133531, doi:10.1016/j.scitotenv.2019.07.337.

Overdevest, C., C. H. Orr, and K. Stepenuck (2004), Volunteer Stream Monitoring and Local Participation in Natural Resource Issues, *Hum. Ecol. Rev.*, *11*(2), 177–185.

Parrish, J. K., H. Burgess, J. F. Weltzin, L. Fortson, A. Wiggins, and B. Simmons (2018), Exposing the Science in Citizen Science: Fitness to Purpose and Intentional Design, *Integr. Comp. Biol.*, *58*(1), 150–160, doi:10.1093/icb/icy032.

Paul, J. D. et al. (2018), Citizen science for hydrological risk reduction and resilience building, *Wiley Interdiscip. Rev. Water*, *5*(1), e1262, doi:10.1002/wat2.1262.

Peckenham, J. M., and S. K. Peckenham (2014), Assessment of quality for middle level and high school student-generated water quality data, *J. Am. Water Resour. Assoc.*, *50*(6), 1477–1487, doi:10.1111/jawr.12213.

Pieper, K. J., R. Martin, M. Tang, L. Walters, J. Parks, S. Roy, C. Devine, and M. A. Edwards (2018), Evaluating Water Lead Levels During the Flint Water Crisis, *Environ. Sci. Technol.*, *52*(15), 8124–8132, doi:10.1021/acs.est.8b00791.

Pool, S., and J. Seibert (2019), Uncertainty guided discharge sampling in ungauged basins: an active-learning approach, in *EGU General Assembly 2019, Geophysical Research Abstracts, Vol. 21, EGU2019-4054*, Vienna.

Power, C., A. Simms, and R. White (2001), Hierarchical fuzzy pattern matching for the regional comparison of land use maps, *Int. J. Geogr. Inf. Sci.*, *15*(1), 77–100, doi:10.1080/136588100750058715.

Price, C. A., and H. S. Lee (2013), Changes in participants' scientific attitudes and epistemological beliefs during an astronomical citizen science project, *J. Res. Sci. Teach.*, *50*(7), 773–801, doi:10.1002/tea.21090.

Reges, H. W., N. Doesken, J. Turner, N. Newman, A. Bergantino, and Z. Schwalbe (2016), CoCoRaHS: The Evolution and Accomplishments of a Volunteer Rain Gauge Network, *Bull. Am. Meteorol. Soc.*, *97*(10), 1831–1846, doi:10.1175/BAMS-D-14-00213.1.

Rey-Mazón, P., H. Keysar, S. Dosemagen, C. D'Ignazio, and D. Blair (2018), Public Lab: Community-Based Approaches to Urban and Environmental Health and Justice, *Sci. Eng. Ethics*, *24*, 971–997, doi:10.1007/s11948-018-0059-8.

Rinderer, M., A. Kollegger, B. M. C. Fischer, M. Stähli, and J. Seibert (2012), Sensing with boots and trousers - qualitative field observations of shallow soil moisture patterns, *Hydrol. Process.*, *26*(26), 4112–4120, doi:10.1002/hyp.9531.

Rinderer, M., H. C. Komakech, D. Müller, G. L. B. Wiesenberg, and J. Seibert (2015), Qualitative soil moisture assessment in semi-arid Africa - the role of experience and training on inter-rater reliability, *Hydrol. Earth Syst. Sci.*, *19*, 3505–3516, doi:10.5194/hess-19-3505-2015.

Rufino, M. C. et al. (2018), Citizen scientists monitor water quantity and quality in Kenya, *CIFOR infobriefs*, (230), doi:10.17528/cifor/007013.

Ruhi, A., M. L. Messager, and J. D. Olden (2018), Tracking the pulse of the Earth's fresh waters, *Nat. Sustain.*, *1*(4), 198–203, doi:10.1038/s41893-018-0047-7.

Sauermann, H., and C. Franzoni (2015), Crowd science user contribution patterns and

their implications, *Proc. Natl. Acad. Sci.*, *112*(3), 679–684, doi:10.1073/pnas.1408907112.

Schrier, K. (2017), Designing Learning with Citizen Science and Games, *Emerg. Learn. Des. J.*, *4*(2017), 19–26.

Science Communication Unit - University of the West of England (2013), *Science for Environment Policy In-Depth Report: Environmental Citizen Science*, Bristol.

See, L., A. Comber, C. Salk, S. Fritz, M. van der Velde, C. Perger, C. Schill, I. McCallum, F. Kraxner, and M. Obersteiner (2013), Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts, *PLoS One*, *8*(7), 1–11, doi:10.1371/journal.pone.0069958.

See, L., T. Sturn, C. Perger, S. Fritz, I. Mccallum, and C. Salk (2014), Cropland Capture : A Gaming Approach to Improve Global Land Cover, in *Connecting a Digital Europe Through Location and Place. Proceedings of the AGILE 2014 International Conference on Geographic Information Science*, edited by Huerta, Schade, and Granell, pp. 3–6, Castellón.

Seibert, J. (2000), Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrol. Earth Syst. Sci.*, *4*(2), 215–224, doi:10.5194/hess-4-215-2000.

Seibert, J. (2001), On the need for benchmarks in hydrological modelling, *Hydrol. Process.*, *15*(6), 1063–1064, doi:10.1002/hyp.446.

Seibert, J., and K. J. Beven (2009), Gauging the ungauged basin: how many discharge measurements are needed?, *Hydrol. Earth Syst. Sci.*, *13*(6), 883–892, doi:10.5194/hess-13-883-2009.

Seibert, J., and J. J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resour. Res.*, *38*(11), 23-1-23–14, doi:10.1029/2001WR000978.

Seibert, J., and M. Vis (2012), Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, *Hydrol. Earth Syst. Sci.*, *16*(9), 3315–3325, doi:10.5194/hess-16-3315-2012.

Seibert, J., and M. J. P. Vis (2016), How informative are stream level observations in different geographic regions?, *Hydrol. Process.*, *30*(14), 2498–2508, doi:10.1002/hyp.10887.

Seibert, J., M. J. P. Vis, E. Lewis, and H. J. van Meerveld (2018), Upper and lower benchmarks in hydrological modelling, *Hydrol. Process.*, *32*(8), 1120–1125, doi:10.1002/hyp.11476.

Seibert, J., H. J. van Meerveld, S. Etter, B. Strobl, R. Assendelft, and P. Hummer (2019), Wasserdaten sammeln mit dem Smartphone – Wie können Menschen messen, was hydrologische Modelle brauchen?, *Hydrol. und Wasserbewirtschaftung*, *63*(2), doi:10.5675/HyWa_2019.2_1.

Silvertown, J. (2009), A new dawn for citizen science, *Trends Ecol. Evol.*, *24*(9), 467–471, doi:10.1016/j.tree.2009.03.017.

Silvertown, J., M. Harvey, R. Greenwood, M. Dodd, J. Rosewell, T. Rebelo, J. Ansine, and K. McConway (2015), Crowdsourcing the identification of organisms: A case-study of iSpot, *Zookeys*, *480*, 125–146, doi:10.3897/zookeys.480.8803.

Soykan, C. U., J. Sauer, J. G. Schuetz, G. S. LeBaron, K. Dale, and G. M. Langham (2016),

Population trends for North American winter birds based on hierarchical models, *Ecosphere*, *7*(5), 1–16, doi:10.1002/ecs2.1351.

Spearman, C. (1904), The Proof and Measurement of Association between Two Things, *Am. J. Psychol.*, *15*(1), 72, doi:10.2307/1412159.

Stahl, K., M. Weiler, I. Kohn, D. Freudiger, J. Seibert, M. Vis, and K. Gerlinger (2016), *The snow and glacier melt components of streamflow of the river Rhine and its tributaries considering the influence of climate change*, Freiburg.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling (2009), eBird: A citizen-based bird observation network in the biological sciences, *Biol. Conserv.*, *142*(10), 2282–2292, doi:10.1016/j.biocon.2009.05.006.

Swanson, A., M. Kosmala, C. Lintott, and C. Packer (2016), A generalized approach for producing, quantifying, and validating citizen science data from wildlife images, *Conserv. Biol.*, *30*(3), 520–531, doi:10.1111/cobi.12695.

The University of North Carolina at Chapel Hill (2019), Lake Observations by Citizen Scientists & Satellites, Available from: www.locss.org (Accessed 14 December 2019)

Thornhill, I., A. Chautard, and S. Loiselle (2018), Monitoring Biological and Chemical Trends in Temperate Still Waters Using Citizen Science, *Water*, *10*(7), 839, doi:10.3390/w10070839.

Thornhill, I., S. Loiselle, W. Clymans, and C. G. E. van Noordwijk (2019), How citizen scientists can enrich freshwater science as contributors, collaborators, and co-creators, *Freshw. Sci.*, *38*(2), 231–235, doi:10.1086/703378.

Turner, D. S., and H. E. Richter (2011), Wet/dry mapping: using citizen scientists to monitor the extent of perennial surface flow in dryland regions., *Environ. Manage.*, *47*(3), 497–505, doi:10.1007/s00267-010-9607-y.

Vörösmarty, C. J. et al. (2001), Global water data: A newly endangered species, *Eos (Washington. DC).*, *82*(5), 1999–2001, doi:10.1029/01EO00031.

de Vries, M., A. Land-Zandstra, and I. Smeets (2019), Citizen Scientists' Preferences for Communication of Scientific Output: A Literature Review, *Citiz. Sci. Theory Pract.*, *4*(1), 1–13, doi:10.5334/cstp.136.

Walker, D., N. Forsythe, G. Parkin, and J. Gowing (2016), Filling the observational void: Scientific value and quantitative validation of hydrometeorological data from a community-based monitoring programme, *J. Hydrol.*, *538*, 713–725, doi:10.1016/j.jhydrol.2016.04.062.

Weeser, B., J. Stenfert Kroese, S. R. Jacobs, N. Njue, Z. Kemboi, A. Ran, M. C. Rufino, and L. Breuer (2018), Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring, *Sci. Total Environ.*, *631–632*, 1590–1599, doi:10.1016/j.scitotenv.2018.03.130.

Weeser, B., S. Jacobs, P. Kraft, M. C. Rufino, and L. Breuer (2019), Rainfall - Runoff Modeling Using Crowdsourced Water Level Data, *Water Resour. Res.*, *55*, 1–16, doi:10.1029/2019WR025248.

West, S., and R. Pateman (2016), Recruiting and Retaining Participants in Citizen Science: What Can Be Learned from the Volunteering Literature?, *Citiz. Sci. Theory Pract.*, *1*(2), 1–10, doi:10.5334/cstp.8.

Westerberg, I., J. L. Guerrero, J. Seibert, K. J. Beven, and S. Halldin (2011), Stage-discharge

uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras, *Hydrol. Process.*, *25*(4), 603–613, doi:10.1002/hyp.7848.

Western, A. W., R. B. Grayson, and J. F. Costelloe (2005), Principles of Hydrological Measurements, in *Encyclopedia of Hydrological Sciences*, edited by M. G. Anderson and J. J. M. McDonnell, p. 3456, John Wiley & Sons, Ltd.

Whitfield, P. H. (2012), Why the Provenance of Data Matters: Assessing "Fitness for Purpose" for Environmental Data, *Can. Water Resour. J.*, *37*(1), 23–36, doi:10.4296/cwrj3701866.

Wiggins, A., G. Newman, R. D. Stevenson, and K. Crowston (2011), Mechanisms for data quality and validation in citizen science, in *Seventh IEEE International Conference on e-Science Workshops*, pp. 14–19, IEEE, Stockholm.

Wilson, N. J., E. Mutter, J. Inkster, and T. Satterfield (2018), Community-Based Monitoring as the practice of Indigenous governance : A case study of Indigenous-led water quality monitoring in the Yukon River Basin, *J. Environ. Manage.*, *210*, 290–298, doi:10.1016/j.jenvman.2018.01.020.

PAPER I

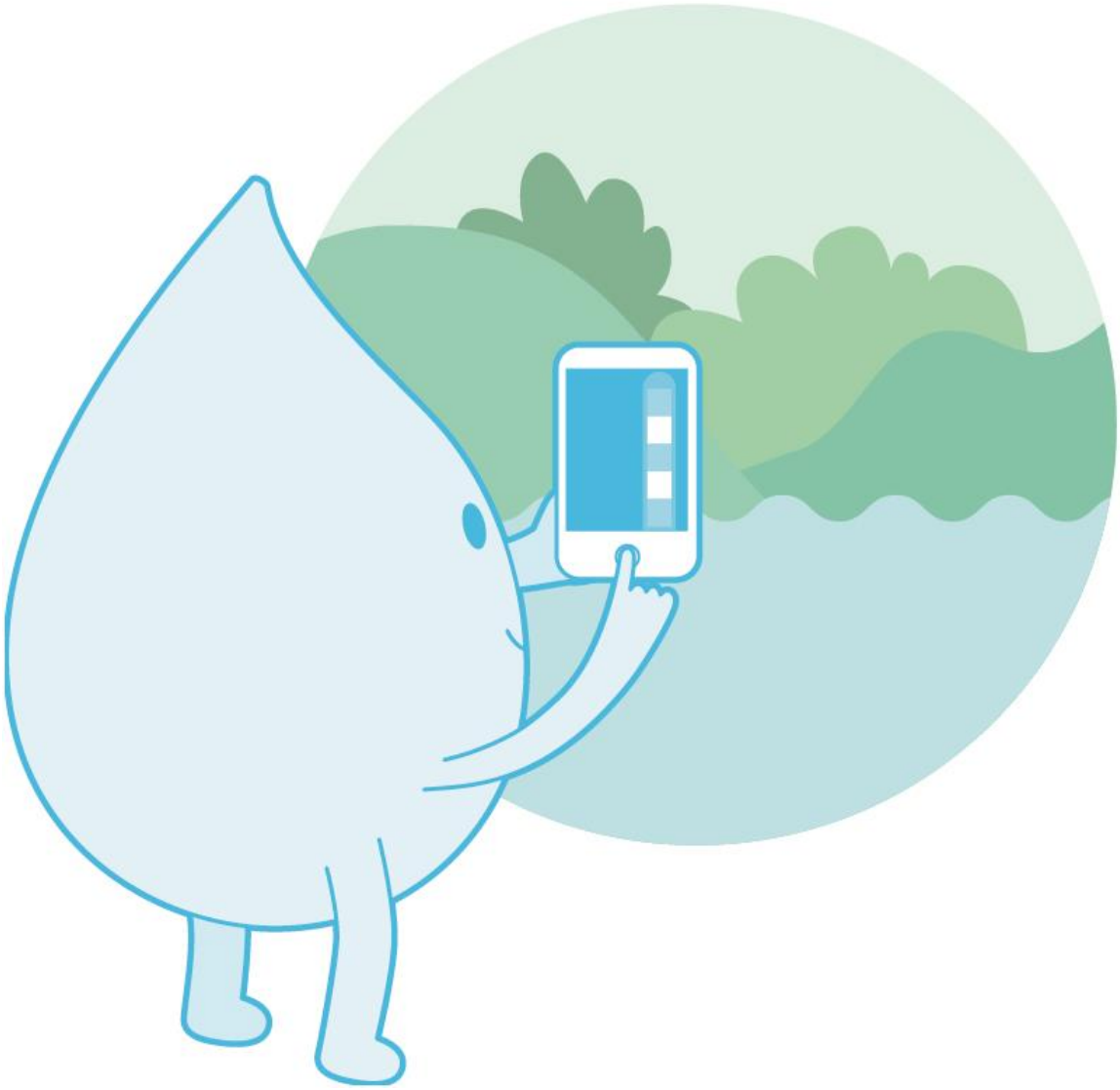


Figure by Tara von Grebel

Seibert, J., **B. Strobl**, S. Etter, P. Hummer, and H.J. van Meerveld (2019), Virtual staff gauges for crowd-based stream level observations, *Frontiers in Earth Science – Hydrosphere*, https://doi.org/10.3389/feart.2019.00070.

frontiers in Earth Science

METHODS
published: 12 April 2019
doi: 10.3389/feart.2019.00070

Virtual Staff Gauges for Crowd-Based Stream Level Observations

Jan Seibert[1,2], Barbara Strobl[1], Simon Etter[1], Philipp Hummer[3] and H. J. (Ilja) van Meerveld[1]*

[1] Department of Geography, University of Zurich, Zurich, Switzerland, [2] Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden, [3] SPOTTERON GmbH, Vienna, Austria

Hydrological observations are crucial for decision making for a wide range of water resource challenges. Citizen science is a potentially useful approach to complement existing observation networks to obtain this data. Previous projects, such as CrowdHydrology, have demonstrated that it is possible to engage the public in contributing hydrological observations. However, hydrological citizen science projects related to streamflow have, so far, been based on the use of different kinds of instruments or installations; in the case of stream level observations, this is usually a staff gauge. While it may be relatively easy to install a staff gauge at a few river sites, the need for a physical installation makes it difficult to scale this type of citizen science approach to a larger number of sites because these gauges cannot be installed everywhere or by everyone. Here, we present a smartphone app that allows collection of stream level information at any place without any physical installation as an alternative approach. The approach is similar to geocaching, with the difference that instead of finding treasure-hunting sites, hydrological measurement sites can be generated by anyone and at any location and these sites can be found by the initiator or other citizen scientists to add another observation at another time. The app is based on a virtual staff gauge approach, where a picture of a staff gauge is digitally inserted into a photo of a stream bank or a bridge pillar, and the stream level during a subsequent field visit to that site is compared to the staff gauge on the first picture. The first experiences with the use of the app by citizen scientists were largely encouraging but also highlight a few challenges and possible improvements.

Keywords: citizen science, smartphone app, water level class, crowdsourcing, data collection

OPEN ACCESS

Edited by:
Jonathan D. Paul, Imperial College London, United Kingdom

Reviewed by:
Wouter Buytaert, Imperial College London, United Kingdom
Tim van Emmerik, Delft University of Technology, Netherlands
Jon Olav Skøien, European Commission – Joint Research Center, Belgium

****Correspondence:***
Jan Seibert
jan.seibert@geo.uzh.ch

Specialty section:
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

Received: *07 September 2018*
Accepted: *19 March 2019*
Published: *12 April 2019*

Citation:
Seibert J, Strobl B, Etter S, Hummer P and van Meerveld HJ (2019) Virtual Staff Gauges for Crowd-Based Stream Level Observations. Front. Earth Sci. 7:70. doi: 10.3389/feart.2019.00070

INTRODUCTION

Data on the quantity and quality of water are needed for appropriate water management decisions. However, hydrology and water resources management are frequently restricted by limited data availability, particularly in data-scarce regions with urgent water management issues (Mulligan, 2013). The decline of national hydrological and meteorological observation networks (Vörösmarty et al., 2001; Fekete et al., 2012; Ruhi et al., 2018) is frustrating, especially in light of the current local and global water-related challenges, and those ahead, such as adaptation to extreme events

and securing water resources for a growing population. Although new observation techniques, including remote sensing, geophysical methods, and wireless sensor networks, provide exciting opportunities for new data collection, central hydrological variables, such as soil moisture or streamflow remain difficult to observe with a sufficient spatiotemporal resolution. Therefore, crowd-based data collection might be a valuable complementary approach to collect data and overcome data limitations (Buytaert et al., 2014).

The idea to include the public in hydrological and meteorological data collection is by no means new. The Swedish meteorologist Tor Bergeron asked the public through appeals over radio and phone calls to measure snow depth (Bergeron, 1949) and rainfall (Bergeron, 1960) and to mail their observations on postcards. This resulted in much more detailed maps than would have been possible with official station data alone. It allowed the creation of a snow depth map for an area of one degree square covering Uppland, Sweden based on 98 observations by volunteers rather than data from only 12 official stations (Bergeron, 1949). For the rainfall observations, Bergeron and his co-workers developed the Pluvius rain gauge as an inexpensive alternative to existing, official gauges. While later there were ∼800 of these gauges in other parts in Sweden, for the initial surveys during 1953 about 150 gauges were distributed in a ∼30 km by ∼30 km area around Uppsala, Sweden (Bergeron, 1960). Both of these projects led to a better understanding of the influence of topography and vegetation on precipitation formation. Even though these early studies were very successful, similar approaches remained rare due to the logistical challenge to transmit and enter the collected data in a common database. However, recent developments in information and communication technology provide exciting new opportunities for citizen-science based approaches using text messages (Lowry and Fienen, 2013; Weeser et al., 2018), websites (e.g., Stream Tracker[1]), apps (e.g., Teacher et al., 2013; Davids et al., 2018; Kampf et al., 2018; Photrack[2]), data mining (Smith et al., 2015; Li et al., 2018) or custom-designed wearable sensors (e.g., Hut et al., 2016; smartfin[3]). However, as stated by Jerad Bales, the Chief scientist for hydrology at the U.S. Geological Survey, "*Crowdsourcing water-information is in its infancy [. . .], and there remain major issues of data quality and sustainability (Lowry and Fienen, 2013). Nevertheless, the use of crowdsourcing to report routine water data, as well as information on floods and droughts, needs to be creatively explored*" (Bales, 2014).

With a large number of contributions from citizens, the CrowdHydrology project[4] (Lowry and Fienen, 2013) has (and still does) successfully demonstrated that it is possible to engage the public in hydrological measurements by asking them to submit stream level observations via text messages. A similar system was implemented in Cithyd[5]. However, these approaches using staff gauges (scaled measurement sticks in the water) restrict the number of places where stream levels can be observed because staff gauges cannot be installed everywhere and by everyone. In mountainous streams, a stable installation is challenging even for hydrologists, and often permits are required before a staff gauge can be installed. Furthermore, if a physical installation is possible, one might consider installing a stream level logger instead of a staff gauge as these loggers have become less expensive and more reliable in recent years. Instead, we propose an approach where anyone can start a measurement location and the observations can be taken anywhere and by anyone. Our approach is similar to geocaching[6], with the difference that instead of treasure hunting sites, stream level observation sites are established and can be revisited by other citizen scientists. In this paper, we describe the virtual staff gauge approach, highlight several design considerations, and discuss whether people understand the concept. In another study (Strobl et al., 2019), we found that most people can classify the water level correctly by comparing it to a reference picture with a virtual staff gauge. Here the focus was on how well people are able to "install" a virtual staff gauge in the app, i.e., taking the reference picture and placing the staff gauge in this picture.

VIRTUAL STAFF GAUGE

General Approach

The advantage of the virtual staff gauge approach is that it avoids physical installations and makes the setup of new observation sites fast and easy. The basic idea behind our approach for stream level observations is that it is usually possible to identify a number of features in a stream or on the streambank, such as rocks, that allow ranking of the stream levels (i.e., "below this tree but above that rock"). While such stream level class observations are not as precise as continuous stream level observations from a staff gauge (i.e., no millimeter resolution) and provide more qualitative information such as "the water level is very low" or "there is a flood event," they can be quite informative for hydrological modeling (van Meerveld et al., 2017). The challenge is to allow easy identification of the different stream level classes, without the need for lengthy verbal descriptions. A picture is helpful in this respect but needs to be amended by a scale. For this, we use the virtual staff gauge approach (see also **Figure 1**):

- The user chooses a suitable site along a stream and identifies the location on a map in the smartphone app.
- The user takes a picture of the streambank (perpendicular to the flow direction and as level as possible, to minimize contortion of the view). There should be some reference in the picture, such as a bridge or stones and ideally, the picture is taken during low flow conditions.
- An image of a yardstick with a number of classes is digitally inserted into the picture as a virtual staff gauge. The user can move the inserted staff gauge in the image and scale it so that it covers the expected stream level variations.

[1]http://www.streamtracker.org

[2]http://www.photrack.ch/mobile.html

[3]https://smartfin.org/

[4]http://www.crowdhydrology.com

[5]http://www.cithyd.com/it/

[6]https://www.geocaching.com/

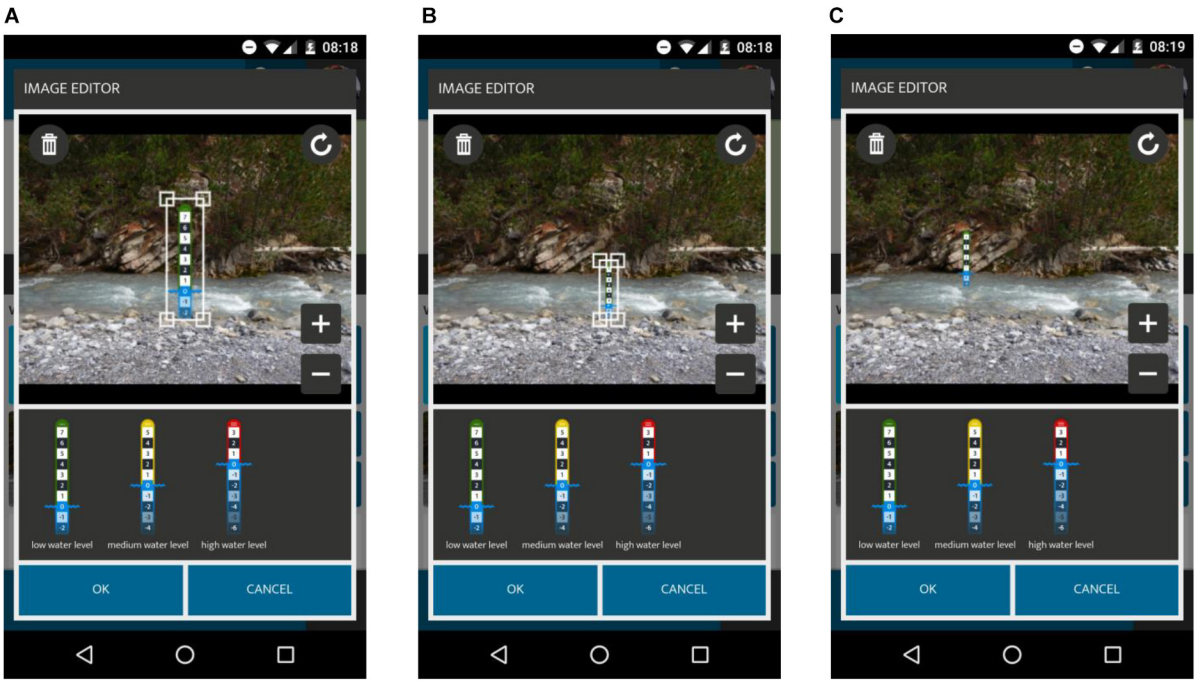


FIGURE 1 | Series of screenshots showing the insertion of the virtual staff gauge in the reference picture: **(A)** insert the image of the staff gauge in the reference picture, **(B)** scale the inserted image, and **(C)** move the image so that the blue line matches the stream level in the picture.

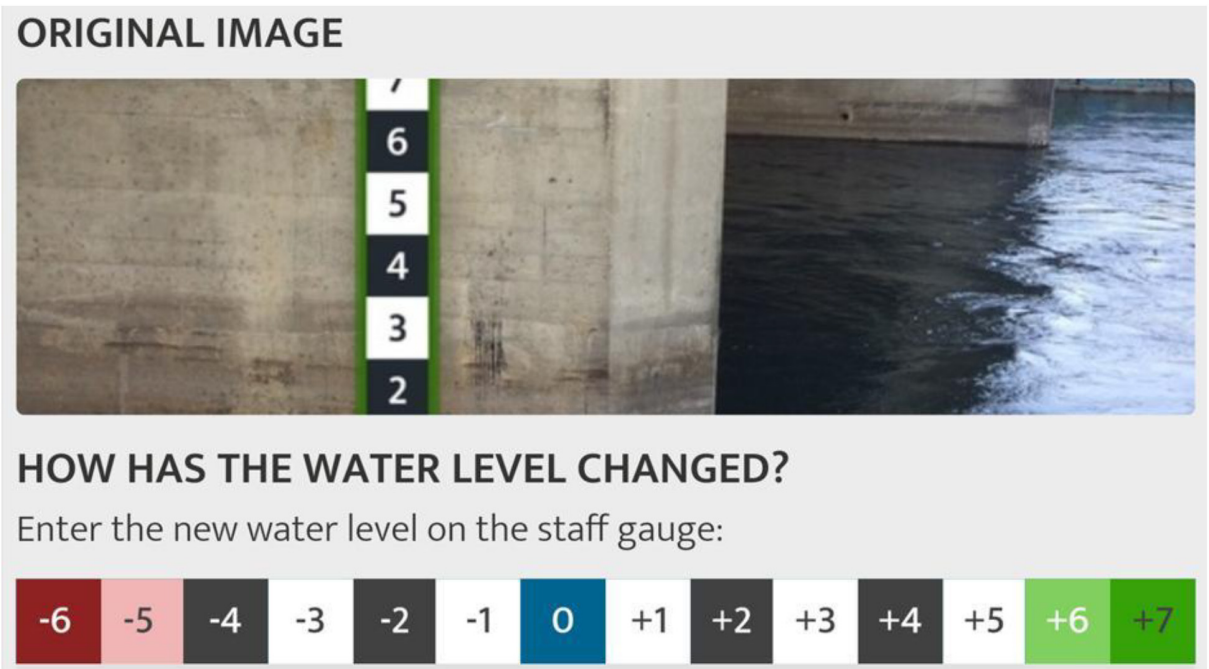


FIGURE 2 | The horizontal version of the staff gauge at the "Update Spot" interface as selectable buttons to report the new water level class observation. Design/author: Philipp Hummer, SPOTTERON Citizen Science, www.spotteron.net.

This reference picture with the virtual staff gauge allows anyone who visits the site at a later time to estimate the stream level class by relating the current stream level to the features on the photo and the virtual staff gauge (e.g., the stream level has changed and is now above a certain rock). For this update, a simplified horizontal staff gauge design is used in the "Update

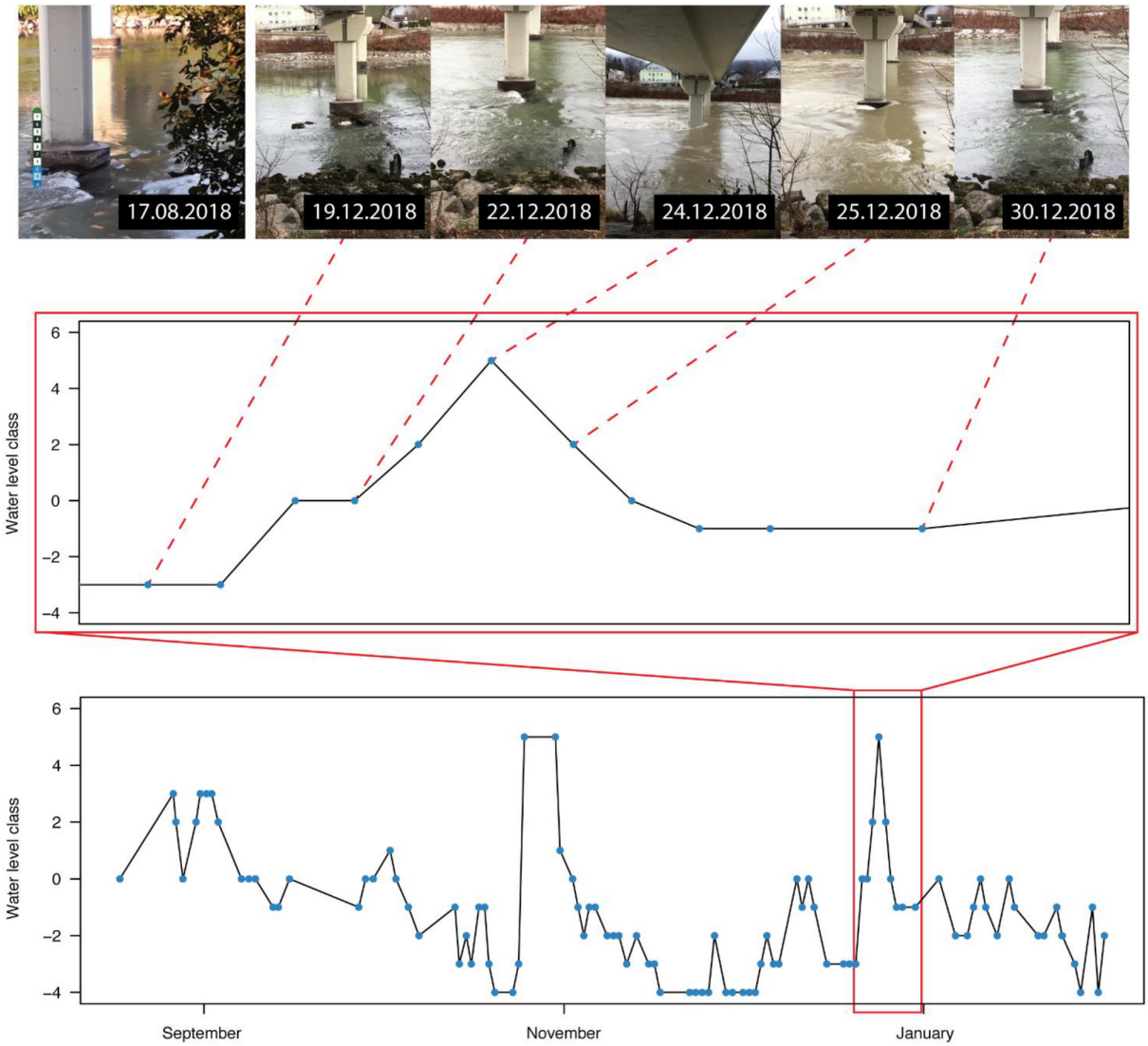


FIGURE 3 | Example of a water level time series obtained using the CrowdWater app (River Salzach, Austria). The pictures for one runoff event (and the reference picture) are shown as an example in the top row.

Spot" interface of the app (**Figure 2**) that shows the full range of class bars for input. To update a spot and provide a new observation of the stream level, the user compares the current stream level with the reference picture with the staff gauge in the app, takes a new picture of the stream, clicks on the current stream level class on the horizontal staff gauge and submits the new observation to the data servers. Over time, this results in a time series of water level observations (**Figure 3**). It is important to note, that the user observes and enters the water level; the new picture is only used for documentation. While automated image recognition could be valuable, at this point we rather rely on human eyes and interpretation and avoid issues such as the exact location and angle when the picture is taken. The pictures, however, allow data quality control. We have recently developed the CrowdWater game as an approach to use these pictures for crowdbased quality control of the water level class data (see "Game"[7]).

Design Considerations and Initial Tests

Several decisions on the design of the virtual staff gauge had to be taken before implementation in the smartphone app. Early on it was decided to use relative stream level classes instead of numeric values in, for instance, centimeters, as there is an obvious limitation in the resolution of stream-level observations that can be achieved with a virtual staff gauge. Translating the virtual staff gauge levels to absolute levels would also make the "virtual installation" much more time consuming as it would require observations of different heights.

[7]https://www.crowdwater.ch

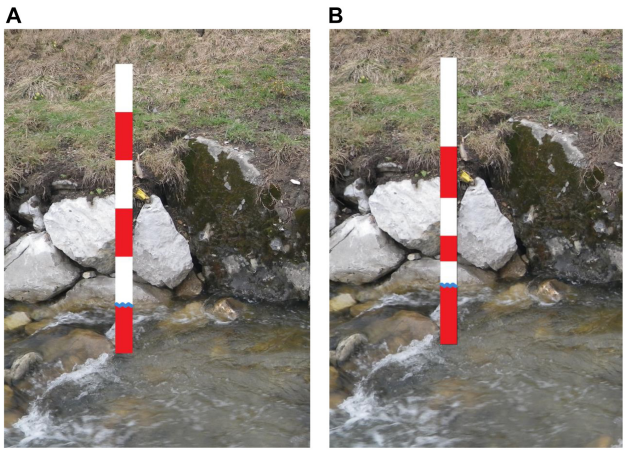


FIGURE 4 | Early version of the virtual staff gauge with regular **(A)** and irregular **(B)** class sizes.

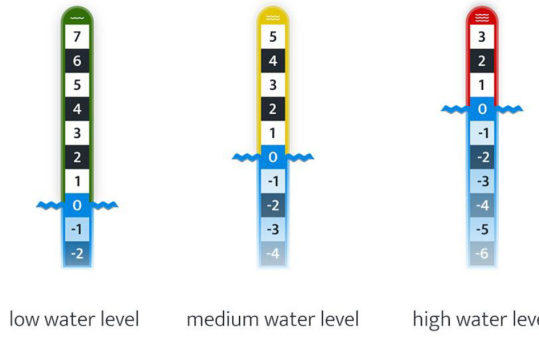


FIGURE 6 | The three staff gauges available in the app. Their ideal application depends on the flow condition of the river at the time that the reference picture is taken. Design/author: Philipp Hummer, SPOTTERON Citizen Science, www.spotteron.net.

Absolute levels would also be site-specific, i.e., the offset would vary largely from place to place. Fortunately, absolute levels are not needed for the potential use in hydrological modeling because the relative values provide important information on the timing of streamflow responses (Seibert and Vis, 2016; van Meerveld et al., 2017).

In an early test with university students, two different types of staff gauges were tested. In addition to regular class sizes (as ultimately implemented in the app), we also tested irregular class sizes (**Figure 4**), but this idea was discarded because some users found it confusing and because it did not allow for as much flexibility as we had hoped.

Once we had decided to have a non-metric virtual staff gauge with regular class sizes, we started to discuss the implementation with SPOTTERON, which is the app company hired to develop the CrowdWater app. During these discussions, the focus was largely on how to make the app intuitive to use. A clearly visible blue wave on the virtual staff gauge was chosen to indicate the stream level at the time that the reference picture was taken (**Figure 5**). During placement, the citizen scientists will highlight the stream level in the photo with the water line in the staff gauge (**Figure 1**). We decided to use ten classes on the virtual staff gauge; this was a compromise between simplicity, resolution, and usability. Through the use of a negative and positive scale, we tried to make the image even more intuitive, as a negative value

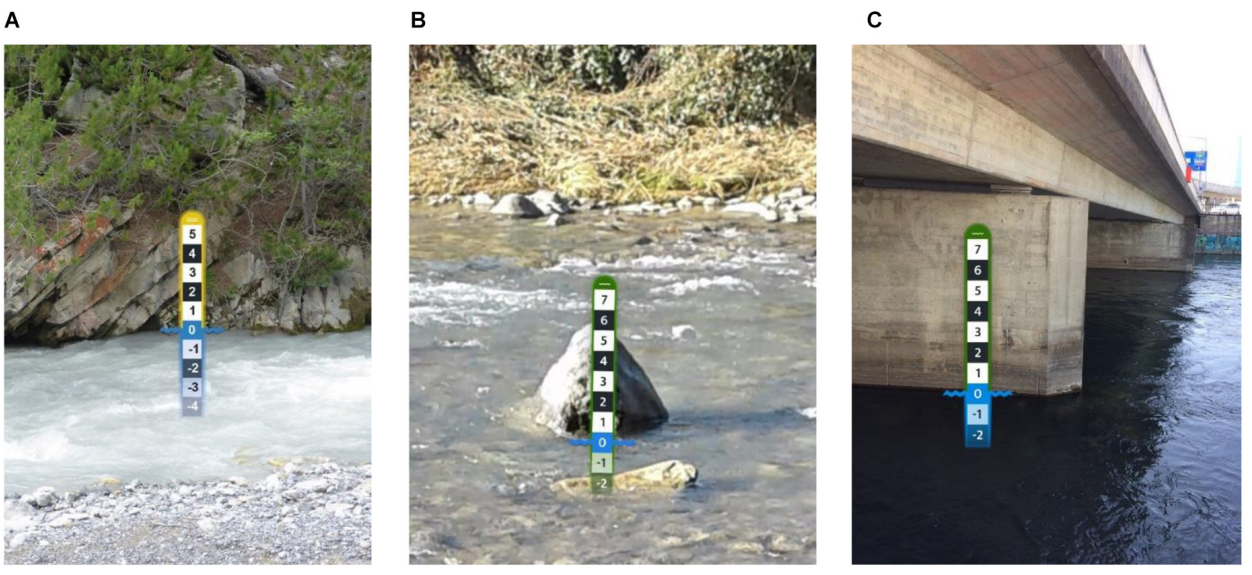


FIGURE 5 | Examples of well-placed virtual staff gauges on **(A)** the opposite stream bank, **(B)** a rock in the stream, and **(C)** a bridge pillar, showing the blue wave at the stream level when the site was established and the positive and negative scale above and below the current stream level, respectively.

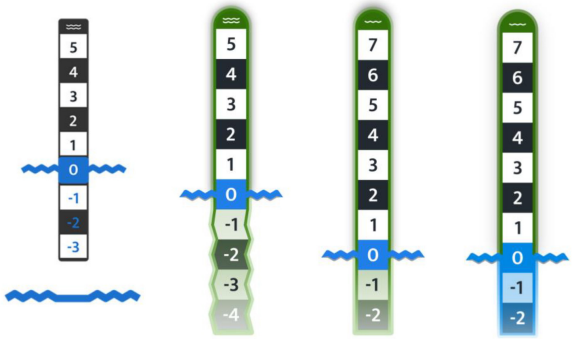


FIGURE 7 | Various staff gauge designs. Design/author: Philipp Hummer, SPOTTERON Citizen Science, www.spotteron.net.

would indicate a stream level below the level in the reference picture and a positive value above it (**Figure 6**). The stream level numbers and class bars follow a neutral black/white scheme to utilize contrast between the sections but also maintain secondary visual weight.

We recommend that citizen scientists initiate a new measurement site during low flow conditions because the reference points are better visible during low flow conditions and this enables future users to better assess the situation for an update. However, this might be a strong restriction in practice and we, therefore, decided to allow insertion of virtual staff gauges also in photos taken during situations with high stream levels. To use suitable staff gauges for all flow conditions, we decided to offer three different staff gauges to the user (**Figure 6**). The green staff gauge is best suited for rivers with a low water level at the time that the reference picture is taken, as it still has many positive classes (i.e., above the blue wave) to record stream levels for higher flow conditions. The yellow staff gauge is well suited for when the reference picture is taken at average flow conditions, and the red staff gauge is ideal for high flow conditions. The red, yellow and green staff gauges were chosen because strong, vibrant colors visually communicate not only a difference but also a development over time, e.g., traffic lights signal different states of movement.

Virtual Staff Gauge Implementation

The virtual staff gauge was implemented as a so-called "sticker". Stickers are a common practice in app design; they use image- or vector-based content as overlays in photos that are taken on a smartphone. They are mainly used in messenger tools, such as WhatsApp or Facebook Messenger to add additional information or emotions to images. Positioning and transformation are usually done by multi-touch gestures for scaling, placement, and rotation. In this case the sticker has to be moved so that the staff gauge is aligned with the streambank or bridge pillar and the blue line is located at the water level (**Figure 1**). By adopting such a rather well-known input method, the use of the app is more intuitive and, thus, optimizes usability. Obviously, using an established technique also had technical advantages for the implementation.

In practice, the placement of the staff gauge can happen on bright or dark, blurry or clear, high- or low-saturation pictures, taken by the users on all kinds of smartphone models and cameras. Therefore, various designs for the virtual staff gauges were tested on different backdrop images and directly on smartphone screens (**Figures 7**, **8**). To ensure that the staff gauge is visible in various conditions, we used additional soft shadows to enhance the edge contrast, but still let the staff gauge immerse itself into the picture as part of the scenery. We furthermore decided to strengthen the visual representation of the areas above and below the stream level by using a blue hue for all class bars below the water level and making them slightly transparent (**Figures 6–8**).

TEST OF THE APP IN PRACTICE

CrowdWater App

The virtual staff gauge was implemented in the CrowdWater smartphone app. The app was first launched for iOS and

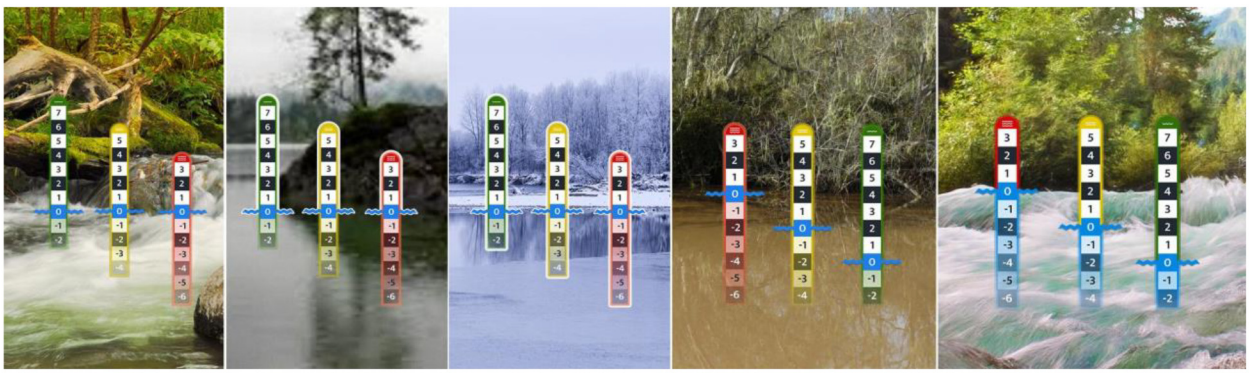


FIGURE 8 | Staff gauge design variants in different environments. Design/author: Philipp Hummer, SPOTTERON Citizen Science, www.spotteron.net. Note that the virtual staff gauges were not scaled nor placed correctly (see **Figure 1**).

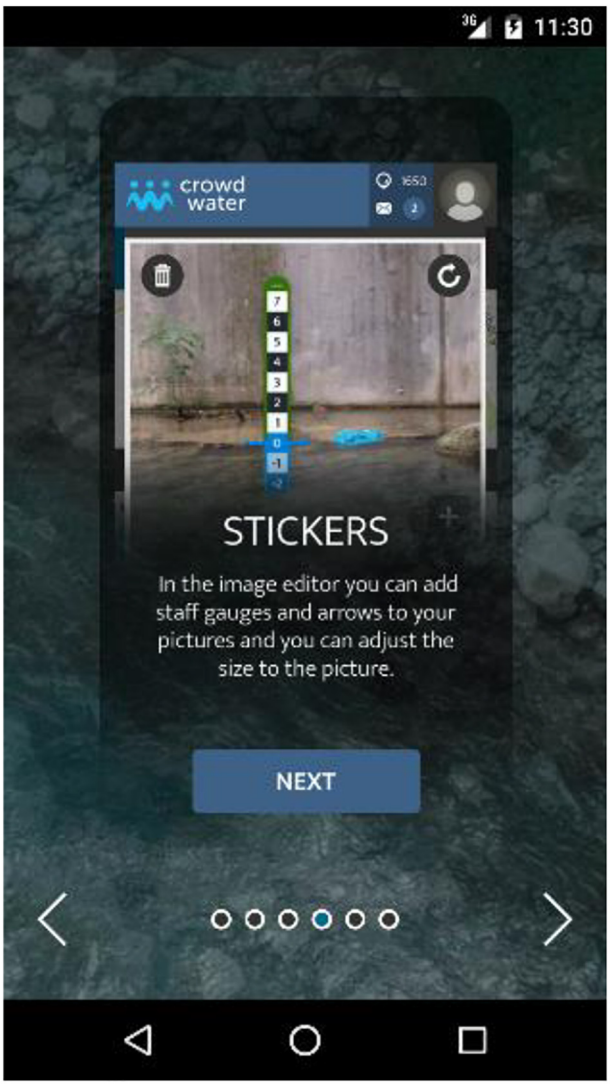


FIGURE 9 | Screenshot of an intro slide that appears when the app is opened for the first time. These can be re-watched anytime. The goal is to quickly provide the most important information on the basic functionalities of the app.

Android in March 2017; there have been several updates of the app since its initial launch. The app was promoted on the CrowdWater homepage (see Footnote 7), through Facebook, Twitter, Instagram, LinkedIn, and ResearchGate posts, as well as on the CrowdWater YouTube channel and at several conferences.

When starting the app, the user has to browse through a number of intro-slides that explain the basic functionalities and the interface of the app. Among them is the sticker function of the virtual staff gauge (**Figure 9**). Additional guidance on how to use the app in the form of texts, pictures and videos are provided on the project homepage and in an explanatory YouTube video[8].

[8]https://www.youtube.com/watch?v=3ag4sHWf0yg

TABLE 1 | Collection of errors made by app-users grouped into broader error categories and frequency of occurrence.

Error type		Frequency of occurrence
Staff gauge size problem	Staff gauge too big	+++
	Staff gauge too small	+
Staff gauge placement problem	Wrong angle	+++
	Staff gauge not on the water surface	+++
Unsuitable location	Lack of reference structure for stream level identification	++
	Structure hidden by vegetation or snow	+
	Unclear which structure to use	+
	River bank too far away	++
	Poor image quality	+
	Site not easily accessible	.
	No suitable site for staff gauge placement available	.
	Changes in the rating curve	+
	Multiple measurement sites at (almost) the same location	+
	Testing (e.g., beer glasses, not a river, out of a train, etc.)	++

+++: occasional = more than 10 times; ++: seldom = 5–10 times; +: rare: less than 5 times; . : not quantifiable.

Typical Mistakes

While users seem to understand the approach used in the CrowdWater app in general, there were also a number of recurrent mistakes related to the staff gauge placement or size. These mistakes affect about 10% of the more than 500 reference pictures (**Table 1**). Staff gauge placement or size problems could be due to users not having read the available instruction material or not fully understanding the concept. Some other issues are not directly related to setting up a virtual staff gauge site but still affect the results, e.g., it is less useful if users create new measurement sites in, or close to, a location where another spot already exists than when they update the existing spot or start a new site on a different river.

Staff Gauge Placement Problem

The most common mistake was related to the placement of the virtual staff gauge. Some users took pictures in the direction of the flow (instead of perpendicular to the flow, see example in **Figure 10**). This makes it almost impossible to place a virtual staff gauge that allows subsequent level observations because clear reference features are usually missing on these pictures. Another placement related issue occurs when the blue wave of the staff gauge is not located at the water surface in the reference picture. This means that the stream level of the reference picture is not at zero, which could lead to confusion for other users when updating the spot later on.

Staff Gauge Size Problems

In a number of cases, the size of the staff gauge was suboptimal. This may be either because people do not realize that they

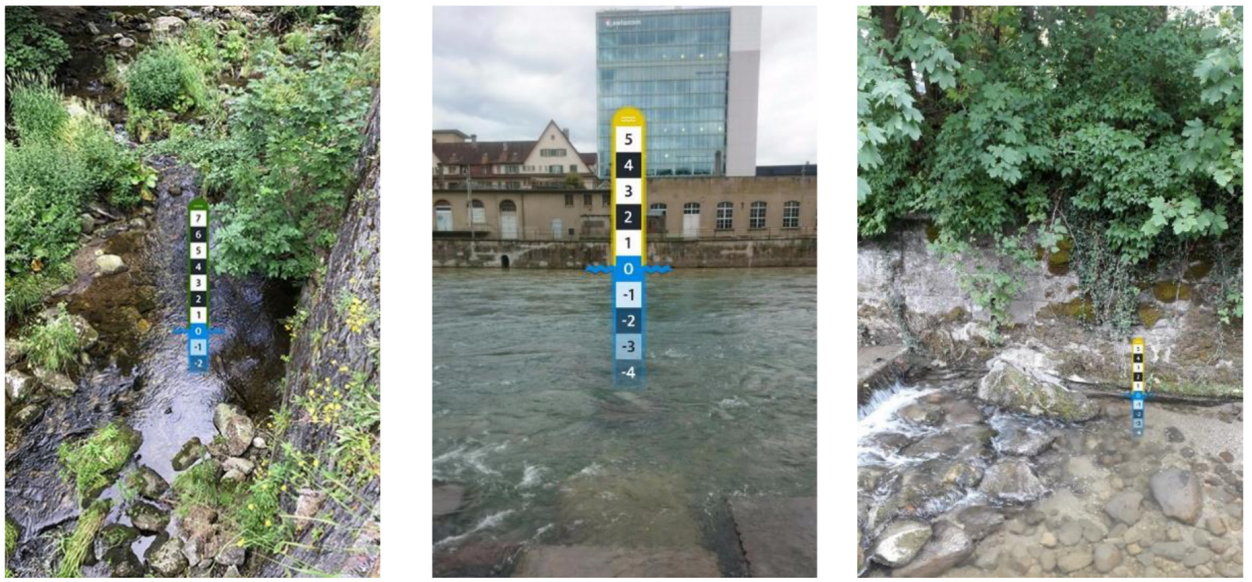


FIGURE 10 | Examples of misplaced virtual staff gauges: **(A)** The picture was taken in the upstream direction instead of perpendicular to the flow direction, which makes it impossible to estimate subsequent stream level changes, **(B)** The virtual staff gauge is so large that it is unlikely that the water level will reach different classes and is therefore improbable to obtain an approximate representation of the stream hydrograph, **(C)** The small virtual staff gauge can show small changes in the stream level, but cannot represent very high flows as anything above a medium flow falls into the highest class.

can resize the size of the staff gauge or do not understand why it is useful to rescale the staff gauge. The perfect staff gauge size is however, somewhat subjective and might to some degree depend on the specific research question and data needs for a site.

In our instruction material, we show the optimal case where the highest class of the staff gauge reaches up to the level of the highest in-bank flow. This may, however, be hard to imagine for citizen scientists and is probably also not considered when users place their first virtual staff gauge. Staff gauges that are too large are not only unrealistic (i.e., the stream level is very unlikely to rise into the highest classes) but this also reduces the variation in future observations because it is less likely that a change in stream level is large enough to reach the next class. There were also a few cases where the staff gauge was too small. A small staff gauge can make it hard to determine the class of the current stream level because the differences between the classes are too small. It also makes it hard to document very high or very low flows. Furthermore, finding the location of the measurement site can be challenging when users take a very detailed (zoomed-in) picture of the reference structure. This issue was more common for small staff gauges and could probably be solved by implementing an option to add an overview photo that shows the general location of the reference structure.

Unsuitable Location

An obvious problem are pictures that lack references for level identification or pictures where a staff gauge was not inserted in the picture. Optimal conditions to place a virtual staff gauge, such as a vertical wall on the opposite river bank or a vertical structure like a rock or bridge pillar in the river, are sometimes hard to find. At least in some cases, the reason for problematic pictures could also be that the rivers were not easily accessible or had no suitable reference features but people still wanted to take a picture to establish a measurement site. Another problem is that in some locations the vegetation growth obscures features on the river bank that were visible when the reference picture was taken (e.g., in winter when there was no vegetation). This makes it nearly impossible to compare stream levels properly. Reference pictures with snow can also make it difficult to assess the stream level later on.

On wide rivers, it is difficult to place a reasonably sized staff gauge at the opposite river bank and still observe changes in stream levels. Furthermore, in these cases, the quality of the pictures is often low due to zooming. This problem can be solved at locations with an instream structure (such as a bridge pillar) and placing the staff gauge along a pillar.

Changes due to erosion or sedimentation are another issue. In these cases stream levels are not a reliable indicator of streamflow. Our dataset contains one site where the riverbed changed quite drastically due to deposited sediment. Because the reference structure (a concrete wall next to a bridge) stayed in place, approximately the same flow meant a different stream level class compared to the situation in the reference picture taken before the sediment was deposited. The solution

to this problem would be to archive the reference picture and create a new one.

CONCLUDING REMARKS

In this paper, we presented a new citizen science approach based on virtual staff gauges that allow crowd-based stream level observations along any stream. The advantage of this approach is that no physical installations are needed, which makes the approach fully scalable, as it is easy and quick for anyone to set up a new measurement site or contribute an observation to an existing site. As discussed in this paper, during development and testing of the virtual staff gauge approach, we identified several issues that required modifications in the original design. Further app developments and better guidance for app users on how to set up a virtual staff gauge site will reduce the number of incorrect sites in the future. Despite these challenges, the first experiences from using the virtual staff gauge approach are encouraging and show that this approach can be useful to collect stream level data at many locations by citizen scientists.

In the first year since launching the smartphone app, numerous measurement sites have been set up. On 3. September 2018, 2431 observations had been submitted by 218 users. For 79 of the 675 sites, more than five updates on the stream level class had been submitted. The collected data have a limited resolution due to the use of stream level classes and are sometimes spotty in time. However, previous work using synthetic data indicates that such data are still informative to constrain hydrological models. Time series of precipitation and temperature are more likely to be available than those of streamflow. The observed stream level class data can, thus, be used in combination with these time series to generate modeled streamflow time series. The potential value of such data has been evaluated based on subsets of existing data. These studies have indicated the value of water level class data for model calibration (van Meerveld et al., 2017); uncertain streamflow estimates were less informative (Etter et al., 2018). The water level data collected in the CrowdWater project are publicly available, and we expect them also to be used for other uses, be it for research, flood protection or leisure activities.

While our current focus is on measurement sites in Switzerland, the app can be, and is already, used worldwide. For developing and evaluating the value of the data obtained with the virtual staff gauge approach countries with a relative wealth of stream data, such as Switzerland, are favorable, but we anticipate that, once developed and tested, the approach will be most beneficial in regions where data are scarce.

AUTHOR CONTRIBUTIONS

JS and HvM developed the first idea of the virtual staff gauge while hiking along a Swiss creek. BS and SE were responsible for the tests and the evaluation of the user experience of the app and contributed by specifying the requirements for the app, which were then discussed among all authors and further developed with PH. PH was responsible for most of the graphical design and the implementation of the smartphone app. JS wrote the manuscript with input from all authors.

FUNDING

The CrowdWater project is funded by the Swiss National Science Foundation (Project Number 163008).

ACKNOWLEDGMENTS

We thank all participants of the CrowdWater project for contributing their observations.

REFERENCES

Bales, J. D. (2014). Progress in data collection and dissemination in water resources - 1974-2014. *Water Resour. Impact* 16, 18–23.

Bergeron, T. (1949). The problem of artificial control of rainfall on the globe. Part II: the coastal orographic maxima of precipitation in autumn and winter. *Tellus* 1, 15–32. doi: 10.1111/j.2153-3490.1949.tb01264.x

Bergeron, T. (1960). *Operation and Results of "Project Pluvius"*. Washington, DC: American Geophysical Union, 152–157. doi: 10.1029/GM005p0152

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., et al. (2014). Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Front. Earth Sci.* 2:26. doi: 10.3389/feart.2014.00026

Davids, J. C., Rutten, M. M., Shah, R. D. T., Shah, D. N., Devkota, N., Izeboud, P., et al. (2018). Quantifying the connections—linkages between land-use and water in the Kathmandu Valley. *Nepal. Environ. Monit. Assess.* 190:17. doi: 10.1007/s10661-018-6687-2

Etter, S., Strobl, B., Seibert, J., and van Meerveld, I. (2018). Value of uncertain streamflow observations for hydrological modelling. *Hydrol. Earth Syst. Sci.* 22, 5243–5257. doi: 10.5194/hess-2018-355

Fekete, B. M., Looser, U., Pietroniro, A., and Robarts, R. D. (2012). Rationale for monitoring discharge on the ground. *J. Hydrometeorol.* 13, 1977–1986. doi: 10.1175/JHM-D-11-0126.1

Hut, R., Tyler, S., and Van Emmerik, T. (2016). Proof of concept: temperature-sensing waders for environmental sciences. *Geosci. Instrum. Methods Data Syst.* 5, 45–51. doi: 10.5194/gi-5-45-2016

Kampf, S., Strobl, B., Hammond, J., Annenberg, A., Etter, S., Martin, C., et al. (2018). Testing the waters: mobile apps for crowdsourced streamflow data. *EOS* 99, 30–34. doi: 10.1029/2018EO096355

Li, Z., Wang, C., Emrich, C. T., and Guo, D. (2018). A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. *Cartogr. Geogr. Inf. Sci.* 45, 97–110. doi: 10.1080/15230406.2016.1271356

Lowry, C. S., and Fienen, M. N. (2013). CrowdHydrology: crowdsourcing hydrologic data and engaging citizen scientists. *Ground Water* 51, 151–156. doi: 10.1111/j.1745-6584.2012.00956.x

Mulligan, M. (2013). WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally. *Hydrol. Res.* 44:748. doi: 10.2166/nh.2012.217

Ruhi, A., Messager, M. L., and Olden, J. D. (2018). Tracking the pulse of the Earth's fresh waters. *Nat. Sustain.* 1, 198–203. doi: 10.1038/s41893-018-0047-7

Seibert, J., and Vis, M. J. P. (2016). How informative are stream level observations in different geographic regions? *Hydrol. Process.* 30, 2498–2508. doi: 10.1002/hyp.10887

Smith, L., Liang, Q., James, P., and Lin, W. (2015). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *J. Flood Risk Manag.* 10, 370–380. doi: 10.1111/jfr3.12154

Strobl, B., Etter, S., van Meerveld, I., and Seibert, J. (2019). Accuracy of crowdsourced streamflow and stream level class estimates. *Hydrol. Sci. J.* (in press). doi: 10.1080/02626667.2019.1578966

Teacher, A. G. F., Griffiths, D. J., Hodgson, D. J., and Inger, R. (2013). Smartphones in ecology and evolution: a guide for the app-rehensive. *Ecol. Evol.* 3, 5268–5278. doi: 10.1002/ece3.888

van Meerveld, H. J., Vis, M. J. P., and Seibert, J. (2017). Information content of stream level class data for hydrological model calibration. *Hydrol. Earth Syst. Sci.* 21, 4895–4905. doi: 10.5194/hess-21-4895-2017

Vörösmarty, C. J., Askew, A., Grabs, W., Barry, R. G., Birkett, C., Döll, P., et al. (2001). Global water data: a newly endangered species. *Eos Trans. Am. Geophys. Union* 82, 1999–2001. doi: 10.1029/01EO00031

Weeser, B., Stenfert Kroese, J., Jacobs, S. R., Njue, N., Kemboi, Z., Ran, A., et al. (2018). Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. *Sci. Total Environ.* 631–632, 1590–1599. doi: 10.1016/j.scitotenv.2018.03.130

Conflict of Interest Statement: PH is founder and co-owner of the company SPOTTERON GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PAPER II

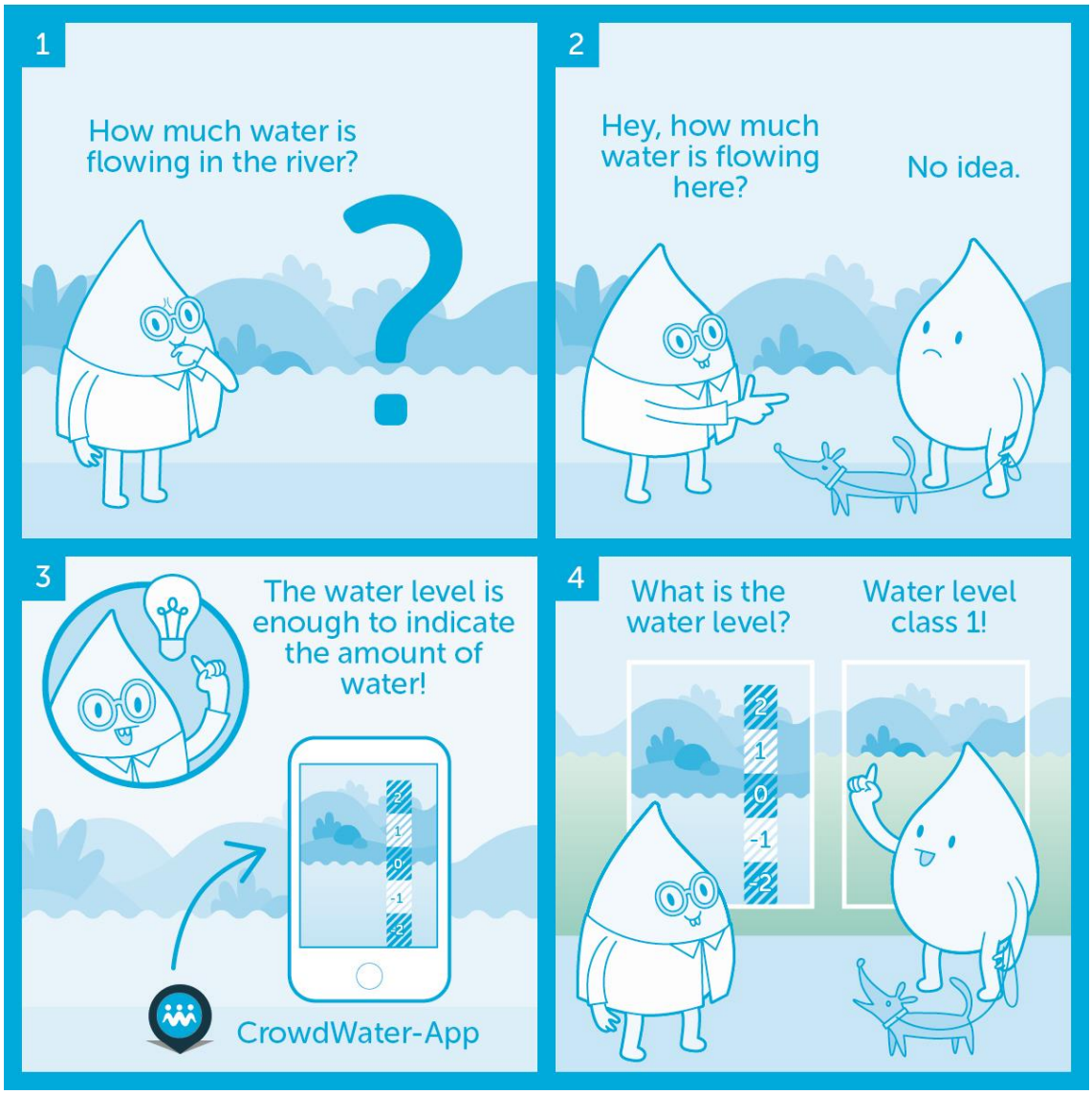


Figure by Tara von Grebel

Strobl, B., S. Etter, H.J. van Meerveld, and J. Seibert (2019), Accuracy of crowdsourced streamflow and stream level class estimates, *Hydrological Sciences Journal, Special Issue: Hydrological Data: Opportunities and Barriers*, https://doi.org/10.1080/02626667.2019.1578966.

HYDROLOGICAL SCIENCES JOURNAL
https://doi.org/10.1080/02626667.2019.1578966

SPECIAL ISSUE: HYDROLOGICAL DATA: OPPORTUNITIES AND BARRIERS

Accuracy of crowdsourced streamflow and stream level class estimates

Barbara Strobl [a], Simon Etter [a], Ilja van Meerveld [a] and Jan Seibert [a,b]

[a]Department of Geography, University of Zurich, Zurich, Switzerland; [b]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

ABSTRACT

Streamflow data are important for river management and the calibration of hydrological models. However, such data are only available for gauged catchments. Citizen science offers an alternative data source, and can be used to estimate streamflow at ungauged sites. We evaluated the accuracy of crowdsourced streamflow estimates for 10 streams in Switzerland by asking citizens to estimate streamflow either directly, or based on the estimated width, depth and velocity of the stream. Additionally, we asked them to estimate the stream level class by comparing the current stream level with a picture that included a virtual staff gauge. To compare the different estimates, the stream level class estimates were converted into streamflow. The results indicate that stream level classes were estimated more accurately than streamflow, and more accurately represented high and low flow conditions. Based on this result, we suggest that citizen science projects focus on stream level class estimates instead of streamflow estimates.

ARTICLE HISTORY
Received 19 June 2018
Accepted 22 November 2018

EDITOR
A. Castellarin

GUEST EDITOR
C. Cudennec

KEYWORDS
citizen science; crowdsourcing; stream level; stream level class; streamflow; accuracy; CrowdWater

1 Introduction

Streamflow data are important for many aspects of river management, including water allocation and the reduction of flood hazards. Streamflow data are also important for the calibration of hydrological models to predict floods and droughts or the impacts of climate change. Most hydrological models need at least a certain amount of data to be properly "tuned" to a particular catchment (Beven 2012).

Three important aspects define the usability of streamflow data: accuracy, spatial coverage and temporal resolution. Conventional streamflow gauging stations can provide detailed information with high accuracy and temporal resolution, but the spatial coverage is limited. While data from gauging stations are considered accurate, the data can still contain substantial errors due to sensor errors, interpolation and extrapolation of the rating curve and cross-section instability (McMillan *et al.* 2012). Typical relative errors for streamflow are ±50–100% for low flows and ±10–20% for medium or high flows (still within the streambank) (McMillan *et al.* 2012). Similar values were derived by Westerberg *et al.* (2011), who mentioned rating curve related errors of −60% to +90% for low flows and ±20% for medium to high flows.

The temporal resolution of gauging stations is often high. However, due to financial and logistic constraints, only a few sites have a gauging station, hence the spatial coverage is limited. Furthermore, these stations may not be installed at representative locations or might miss certain types of catchments, especially small headwater streams (Kirchner 2006, Bishop *et al.* 2008). Also relatively few measurement stations are located in developing countries. Thus, for many catchments there are no streamflow data available for water management decisions or model calibration.

Although new wireless sensor network technology provides the possibility to expand the measurement networks, the reality is that, due to budget cuts, observation networks often shrink rather than expand (Kundzewicz 1997, Ruhi *et al.* 2018). For example, Ruhi *et al.* (2018) showed that between 1947 and 2016 the number of streamgauges in river basins in the USA decreased by 21%.

Several studies have focused on the minimum number of measurements required to properly calibrate a hydrological model (Perrin *et al.* 2007, Juston *et al.* 2009, Seibert and Beven 2009, Seibert and McDonnell 2015, Vis *et al.* 2015) and have shown that even a few streamflow measurements can vastly improve the performance of a model (Pool *et al.* 2017). While employees of agencies responsible for national or regional gauging station networks could perhaps take a limited number of additional measurements at a few ungauged streams, it is impossible for them to take measurements

CONTACT Barbara Strobl barbara.strobl@geo.uzh.ch
Supplemental data for this article can be accessed here.

at all ungauged streams. An interesting alternative to obtaining streamflow data for more streams is to ask citizen scientists or citizen observers to collect streamflow data.

Citizen science has been used in numerous environmental studies to obtain data with a much higher spatial resolution than is otherwise possible (Dickinson *et al.* 2010, Tulloch *et al.* 2013, Aceves-Bueno *et al.* 2017, Hadj-Hammou *et al.* 2017) and has been used to obtain hydrological data as well (Buytaert *et al.* 2014). For example, citizen science data have been used to fill in spatial and temporal gaps in water quality and stream level data series (Lowry and Fienen 2013, Hadj-Hammou *et al.* 2017) and to obtain groundwater level data across large areas (Little *et al.* 2016). Citizen science could therefore be a complementary approach to collect the stream level and streamflow data that are needed for hydrological model calibration, particularly for the many streams that are currently ungauged. In order to involve as many citizens in data collection as possible and to obtain data for remote areas, approaches are needed to collect these data with very little time and effort and without special equipment.

Despite their potential to complement existing data sources, citizen science data are not without challenges; in particular, the accuracy of crowdsourced data is often discussed (Engel and Voshell 2002, Haklay 2010, See *et al.* 2013, Aceves-Bueno *et al.* 2017). Several studies have examined the accuracy of crowdsourced hydrological data (Turner and Richter 2011, Rinderer *et al.* 2012, 2015, Lowry and Fienen 2013, Peckenham and Peckenham 2014, Breuer *et al.* 2015, Le Coz *et al.* 2016, Little *et al.* 2016, Weeser *et al.* 2018). Lowry and Fienen (2013) found promising results in terms of the accuracy of stream level data from participants who read the level from a staff gauge in a stream close to a hiking path. The root mean square error (RMSE) of the crowdsourced stream level data was approximately 5 mm, which was almost as good as that of pressure transducer data. They concluded that the level of accuracy "*is encouraging since no training was given to the citizen scientists*" (Lowry and Fienen 2013, p. 155). In a similar study by Weeser *et al.* (2018) in Kenya, data collected by citizens were comparable to those of conventional data loggers, although they had a low temporal resolution. Little *et al.* (2016) provided volunteers with equipment to measure the water level in their own wells. They found that the absolute difference of the well readings ranged from 2 to 11 mm and concluded that "*community-based groundwater monitoring provides an effective and affordable tool for sustainable water resources management*" (Little *et al.* 2016, p. 317). Peckenham and Peckenham (2014) analysed groundwater quality data collected by students and concluded that the accuracy varied, but "*it is possible to make precise and accurate measurements consistent with the methods specifications*" (Peckenham and Peckenham 2014, p. 1477).

However, these previous hydrological citizen science studies are not easily scalable to many sites because they require the installation of staff gauges or other instrumentation. Therefore, it is useful to also develop and test citizen science approaches to collect streamflow or stream level data that do not require equipment or the installation of staff gauges, but these new citizen science tasks should be designed "*with the skill of the citizens in mind*" (Aceves-Bueno *et al.* 2017, p. 287). It is likely that many citizens who frequently pass by streams notice high and low flows throughout the seasons. These frequently visited locations could be turned into locations for streamflow or stream level class observations if citizens can accurately estimate streamflow or stream level classes.

Testing the accuracy of citizen science data before starting a citizen science project is crucial for every citizen science project. This ensures that the data collected are sufficiently accurate for the purpose of the project and avoids unnecessarily burdening citizens with tasks that result in data that are in hindsight of limited value due to data accuracy issues. The objective of this study was, therefore, to determine what types of parameters related to streamflow citizens can estimate accurately. We asked 517 citizens to estimate both the streamflow and stream level class and assessed whether one can be estimated more accurately than the other by calculating the corresponding streamflow for each stream level class estimate. Accuracy is defined here as the difference between the estimated value and the measured value, as well as the frequency of extreme outliers. The specific research questions for this study were:

(1) How well can stream level class, streamflow and the different factors of streamflow (width, depth, flow velocity) be estimated by citizens?
(2) To what extent do stream size and flow conditions affect the accuracy of the crowdsourced data?

2 Methodology

2.1 Basic approach and study sites

We conducted 16 field surveys where we asked people to estimate the streamflow, as well as the average width, depth and velocity of the stream, and the stream level class. For the surveys, we selected 10 locations (Table 1; see also Supplementary material, Fig. S1) where we

Table 1. Information on the streams where the field surveys took place. Size classes XS: ≤1 m^3/s; S: >1–50 m^3/s, M: >50–200 m^3/s and L: >200 m^3/s. A map with the survey locations is given in the Supplementary material (Fig. S1). Survey dates given as dd.mm.yyyy.

Stream	Size		Date of survey	No. of participants, *n*	Streamflow (m^3/s)	Source for measured streamflow*	Approx. distance to virtual staff gauge (m)	Comments
Chriesbach (Zurich)	XS		29.09.2017	30	0.38	Salt dilution	5	BSc students: no direct streamflow estimates
Hornbach (Zurich)	XS		19.02.2017	33	0.134	Salt dilution	8	
Irchel (Zurich)	XS		11.03.2017	25	0.01	Salt dilution	1	
Glatt (Zurich)	S		29.09.2017	31	2.8	WWEA, station: 533	11	BSc students: no direct streamflow estimates
Magliasina (Magliaso)	S		28.04.2017	40	16	FOEN, station: 2461	14	High-school students: no stream level class estimates
Schanzen-graben (Zurich)	S		01.04.2017	31	2.6	Salt dilution	16	
Sihl (Zurich)	S	1	18.02.2017	33	7	FOEN, station: 2176	32	Low flow
		2	26.07.2017	31	28			High flow
Töss (Winterthur)	S		12.03.2017	35	9	WWEA, stations: 518, 520 and 581	29	Interpolation between three nearby stations for reference value
Limmat (Zurich)	M	1	29.10.2016	38	59	FOEN, station: 2099	7	No stream level class estimates
		2	08.04.2017	27	83			
		3	02.06.2017	31	107			
		4	09.07.2017	44	75			PhD students Low flow
		5	13.11.2017	31	222			High flow
Aare (Brugg)	L	1	07.01.2017	27	108	FOEN, station: 2016	53	Low flow
		2	10.05.2017	30	389			High flow

* The measured streamflow data were obtained from the Federal Office of the Environment (FOEN; http://hydrodaten.admin.ch/), the Office of Waste, Water, Energy and Air of Canton Zurich (WWEA; www.hydrometrie.zh.ch/) or by salt dilution gauging (Salt dilution).

expected enough people to pass by and have time for the survey. We divided the streams into four different size classes (XS, S, M, L) based on the mean annual streamflow, and, when long-term time series were not available, based on the available measurements:

- XS (Chriesbach, Hornbach and Irchel): ≤1 m^3/s,
- S (Glatt, Magliasina, Schanzengraben, Sihl and Töss): >1–50 m^3/s,
- M (Limmat): >50–200 m^3/s, and
- L (Aare): >200 m^3/s.

To analyse whether the flow conditions affect the accuracy of the estimates, surveys were conducted under high and low flow conditions for three streams: Aare (L), Limmat (M) and Sihl (S).

The aim of the surveys was to get a sufficient number of streamflow estimates for a specific stream on a specific day (our aim was 30 participants per survey to assure statistical significance; Field *et al.* 2013). We therefore used a logistically simple sampling strategy, whereby we personally approached passers-by (similar to Breuer *et al.* 2015) and asked if they would complete the 5-minute survey (i.e., we did not use a targeted approach to capture responses of a representative group of citizens). No data were collected on the percentage of passers-by who participated, but we estimate that about every third person we approached agreed to participate in our survey. In addition, we asked high-school (Magliasina) and university students (Chriesbach, Glatt and Limmat) to fill out the survey during excursions. All surveys took place between October 2016 and September 2017. In total, we received 517 complete surveys: 372 passers-by, 61 participants from a university geography bachelor student excursion (Glatt and Chriesbach), 40 from a high-school student excursion (Magliasina) and 44 from a summer school for PhD students from fields ranging from physics to social sciences (Limmat) (see Table 1). During the group excursions we emphasized the need for individual estimates and limited discussions between the students for the duration of the survey.

The age distribution of all 517 participants corresponds to that of the inhabitants of Zurich (where most field surveys were conducted), although there were fewer participants over the age of 60 (13% of the participants *vs* 19% of the population in Zurich; see Supplementary material, Fig. S2(c) and (d)) (Statistik Stadt Zürich 2017). Also a large number of participants were university educated, roughly 48% compared to 16% of the population in Zurich (Fig. S2(b)) (Statistik Stadt Zürich 2017). There was an almost equal split between male and female participants (Fig. S2(a)).

2.2 Streamflow estimation

Participants were first asked to estimate the streamflow directly. For this direct estimate, we asked them to estimate the flow in m^3/s, or in L/s for the very small streams (XS). This directly estimated streamflow value is referred to as Q_{direct}. This task, understandably, proved to be difficult for some participants because streamflow quantification was difficult and they were unfamiliar with the units. A few participants refused to answer this question, even with a bit of prompting. Some decided to guess, even though they thought it was unlikely to be a realistic value and others deduced on their own that they could estimate the width, mean depth and flow velocity to get an approximate value.

After this initial guess of the streamflow, we explained to the participants that it is possible to estimate the individual factors (width, mean depth and flow velocity) and to derive the streamflow by multiplying these values (Equation (1)). The participants were then asked to estimate the average width, mean depth and velocity of the stream. We also asked them to classify the streambed material. Equation (1) was used to calculate the streamflow using these factors:

$$Q_{factor} = w \cdot d \cdot v \cdot k \quad (1)$$

where Q_{factor} is the estimated streamflow (m^3/s), w is the estimated width (m), d is the estimated mean depth (m), v is the estimated surface flow velocity (m/s) and k is the correction factor to obtain the average velocity from the surface velocity. While some participants still found the quantification difficult, they were more familiar with these units, compared to m^3/s or L/s. Often a value of 0.85 is used for the correction factor k (Welber *et al.* 2016); but it can also be estimated using the logarithmic velocity distribution (Prandtl-von Kármán equation) for turbulent flow based on the surface flow velocity, grain size and stream depth (Dingman 2015). This calculated factor for the mean flow velocity varied for the different estimates of the participants (even for the same stream). For two-thirds of all estimates, the calculated velocity factor was not within the typical range of 0.71–0.95 (Welber *et al.* 2016) due to an unrealistic ratio between the estimated average water depth and estimated streambed roughness. Values lower than 0.71 were adjusted to 0.71 (52% of estimates) and values over 0.95 were adjusted to 0.95 (1% of estimates). When no estimate for streambed roughness was available (this happened only occasionally, except for the entire field survey at Magliasina), the typical velocity correction factor of 0.85 was used (including the participants at Magliasina this corresponds to 13% of all estimates). During the university excursion at the Glatt and Chriesbach, we did not ask for direct stream estimates because most geography bachelor students would likely have applied the indirect estimation method (Q_{factor}) because of lectures on streamflow during their education.

To assess the accuracy of crowdsourced streamflow data, the streamflow estimates were compared to measured streamflow data. Streamflow was measured before or after the surveys (Chriesbach, Hornbach, Irchel and Schanzengraben) or obtained from official gauging station data when these were located near the survey location (Aare, Limmat, Magliasina and Sihl, stations of the Swiss Federal Office for the Environment (FOEN); Glatt and Töss, stations of the Office of Waste, Water, Energy and Air of Canton Zurich (WWEA)) (see Table 1). The methods for the reference measurements for width, mean depth and flow velocity depended on the size and accessibility of the river. These measurements included direct measurements for width and depth with measurement tapes, data on the stream cross-section from FOEN for width and depth (when available), an estimate of the width of the river from Google Maps for wide rivers (Aare and Limmat) and the stick method for flow velocity. Even though these measurements are likely also affected by errors, they were assumed to be the "true" data to which the citizen science estimates could be compared. We assumed that the uncertainty for the measured values is 10% for streamflow (Pelletier 1988), 0.5% for width and 1–3% for depth (Herschy 1971) and roughly 10% for flow velocity (based on our own measurements).

2.3 Stream level class estimation

We also asked participants to estimate the stream level class. Stream level refers to the height of the water in a stream. A stream level class means that this height is expressed on a discrete scale of classes, rather than on a continuous scale. Stream level class data only provide information about whether the stream level is higher or lower than previously, but earlier studies have shown that stream level class data are useful for hydrological model calibration (van Meerveld *et al.* 2017). Thus, the participants were not asked to estimate the stream level in centimetres but to estimate the stream level class. The participants compared the current stream level with a photo of the same stream (taken at an earlier time) with a digitally inserted staff gauge with 10 level classes (Fig. 1, also Supplementary material, Section S2). The staff gauge was scaled so

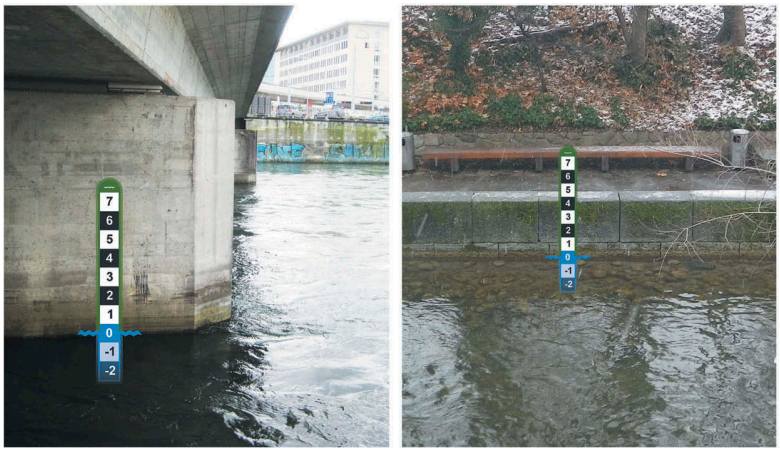


Figure 1. Example of a virtual staff gauge in the pictures used for the surveys at Limmat (left) and Schanzengraben (right). Photographs taken on 29.06.2016 when the streamflow was 165 m^3/s (Limmat) and on 05.01.2017 (unknown streamflow; Schanzengraben). For the dates and the flow conditions during the surveys see Table 1.

that the highest class represented the highest in bank flood level and the lowest class represented the likely lowest stream level. The height of the classes is arbitrary and varied for each location, depending on the size of the river and how the virtual staff gauge was placed in the picture. A small staff gauge would have a higher resolution, but the stream level for very high and low flows may be above or below the staff gauge, whereas a large staff gauge would imply a lower resolution of the observations as the stream level would fluctuate across fewer classes. In this study we tried to place the staff gauges so that the staff gauge covered both high and low in bank flows. The number of classes was a compromise between resolution and usability. A larger number of classes provides higher resolution data but also makes it more difficult (or even impossible) for participants to determine the stream level class. Based on a previous model, study model calibration results do not improve much when more than five stream level classes are used (van Meerveld *et al.* 2017). The number of 10 classes was chosen to ensure observable stream level fluctuations even in cases where the virtual staff gauge is placed so that some classes are never or very rarely reached. The correct stream level class value was determined by us by carefully choosing appropriate references and individually (but unanimously) deciding on the correct stream level class.

For the Limmat, results are given for all five field surveys for streamflow, but stream level class estimates are given for only four surveys because a slightly different virtual staff gauge was used for the first survey.

2.4 Data analyses

To be able to compare the accuracy of the streamflow estimates for different streams, relative estimates (in percent) were calculated by dividing the streamflow estimate by the measured value (i.e., considered true value). A value of 100% corresponds to a perfect estimate, smaller values represent an underestimation and larger values represent an overestimation. The quality of the data was then assessed by statistical measures, such as the interquartile range and median. In addition, we determined the number of outliers as they are likely disinformative for model calibration (Beven and Westerberg 2011) and can be worse than having no data. Even though filters can be used to remove outliers in citizen science data, in practice, it may be difficult to filter out all outliers. All relative estimates below 50% and above 150% were considered to be outliers.

For comparison between streamflow and stream level class estimates, stream level classes and the errors in this classification were converted to an equivalent streamflow (m^3/s), named Q_{level} in the remainder of the manuscript. For the stream locations with a nearby FOEN gauging station (Sihl, Limmat, Aare), the classes of the virtual staff gauge were converted to a metric value by determining the stream depth that corresponded to each stream level class (i.e., mid-point and upper and lower stream level for each class) and using the FOEN rating curve to convert these stream levels to a streamflow estimate. For the sites where no rating curve was available (Hornbach, Irchel, Schanzengraben and Töss), additional measurements of the stream profile and water

surface slope (estimated based on the slope of the streambed) were used to estimate the streamflow for each stream level class using the Manning-Strickler formula (Manning 1891). This curve was fitted to the streamflow measured on the day of the surveys by adjusting the roughness coefficient within predefined boundaries based on the streambed material. The roughness coefficient used for the Manning-Strickler formula introduces some subjectivity and thereby likely increases the uncertainty of the conversion of the stream level class to streamflow compared to FOEN rating curve measurements. Since the stream level classes represent a range of values rather than just one value, the streamflow was not only calculated for the centre value of the level class, but also the class boundaries to obtain the possible range of streamflow values. The estimates from Chriesbach, Glatt and Magliasina were excluded from this analysis (101 of the 517 estimates) because the relevant data were not collected at the time of the surveys.

The differences in the median relative estimates for the different stream size classes were tested for significance using the Kruskal-Wallis test with the *post hoc* procedure based on Dunn (1964). Differences in the median relative streamflow estimates between high and low flow conditions were tested for significance using the Mann-Whitney test. A p-value of 0.05 was used for all statistical tests, unless otherwise indicated.

3 Results

3.1 Streamflow estimates

Although there was a large spread in the streamflow estimates, the median values were surprisingly close to the measured streamflow (Figs 2 and 3). Across all surveys the median of the direct streamflow estimates (Q_{direct}) was closer to the measured value than the estimate based on the factors (Q_{factor}) (median relative estimates of 93 and 80%, respectively, when all surveys were analysed together). However, the interquartile range was smaller for the streamflow calculated from the estimated factors (the first and third quartiles were, respectively, 26 and 309% for Q_{direct} and 39 and 172% for Q_{factor}; Fig. 3), meaning that the streamflow estimates were closer to the measured value for the estimates based on the factors.

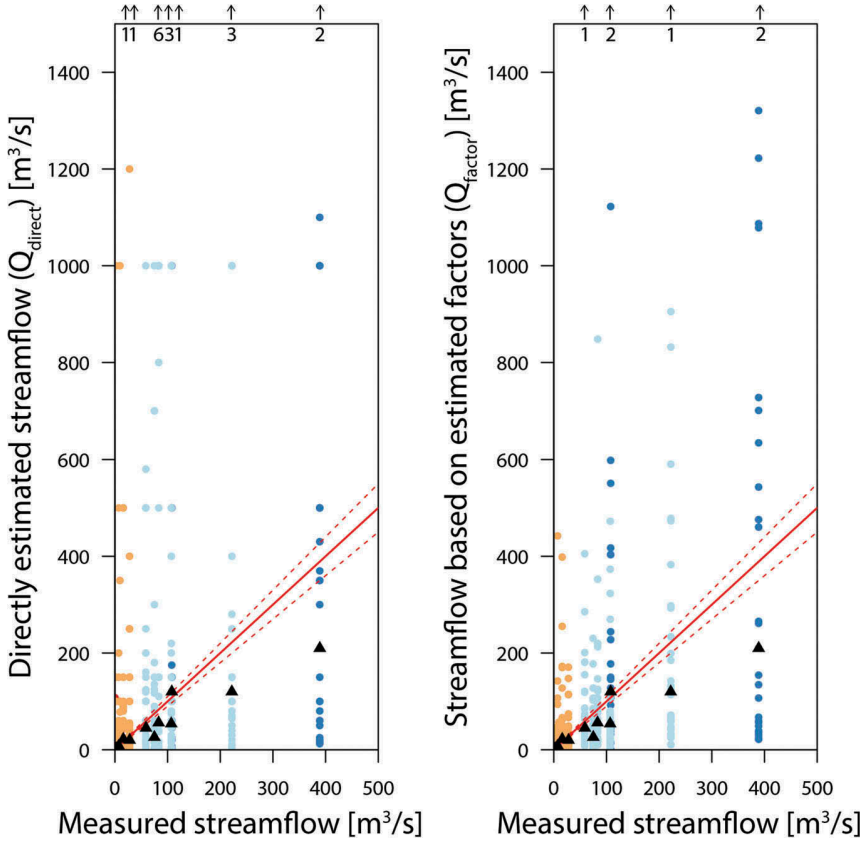


Figure 2. Scatter plots showing the spread of Q_{direct} (left) and Q_{factor} (right) for each field survey. The data points are colour-coded according to the stream size: from left to right, XS to L are red, orange, light blue and dark blue, respectively. ▲: median estimated streamflow per survey; solid and dashed (red) line: the 1:1 line with the 10% uncertainty band. The number at the top of the graph indicates the number of extreme outliers (1–6, not shown).

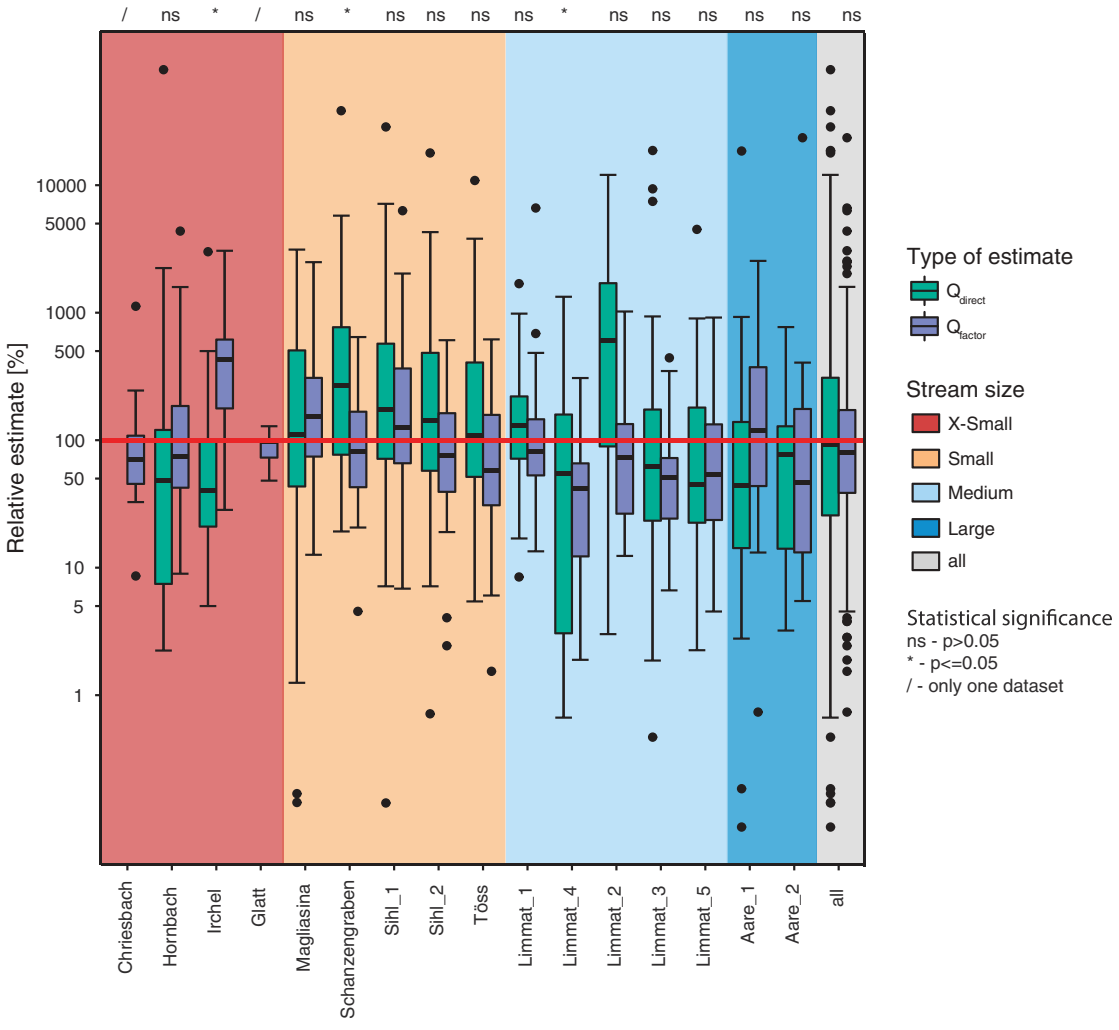


Figure 3. Box plots of the relative estimates of streamflow (ratio of estimated *vs* measured streamflow) for Q_{direct} and Q_{factor} for each survey, and for all streams combined (all). Statistical significance, i.e. difference in median relative streamflow estimate for the two methods, is shown across the top. The data for the Sihl, Limmat and Aare are ordered from low to high flow conditions (see Table 1). The box represents the interquartile range, the black line the median, the whiskers extend to 1.5-times the interquartile range below/above the first/third quartile, and the dots represent values beyond 1.5-times the interquartile range. Note the log scale.

The differences between the median estimates of Q_{direct} and Q_{factor} were statistically significant ($p < 0.05$) for three out of the 14 surveys with both Q_{direct} and Q_{factor} estimates, but not for all surveys combined (Fig. 3). Of these three surveys, two had a median estimate for Q_{direct} that was closer to the measured value. The interquartile range was smaller for Q_{factor} for two of the three surveys.

3.2 Streamflow factor estimates

There were also numerous outliers for the relative estimates of width, mean depth and flow velocity (Fig. 4). The median relative estimates for the width, depth and flow velocity were all significantly different from each other (Fig. 4). The width was generally underestimated (median relative estimate of 75%, and third quartile of 95% when all stream surveys were analysed together), the mean depth was generally overestimated (median relative estimate of 126% when all stream surveys were analysed together), while the median flow velocity was surprisingly accurate (median relative estimate of 100% when all stream surveys were analysed together). However, the interquartile range suggests that width can be estimated most accurately (interquartile range of relative estimates from 57 to 95% when looking at all surveys together), and mean depth (interquartile range of relative estimates from 86 to 180%) and flow velocity (interquartile range of relative estimates from 57 to 143%) can be estimated less accurately. The percentage of relative estimates below 50% or above 150% shows the

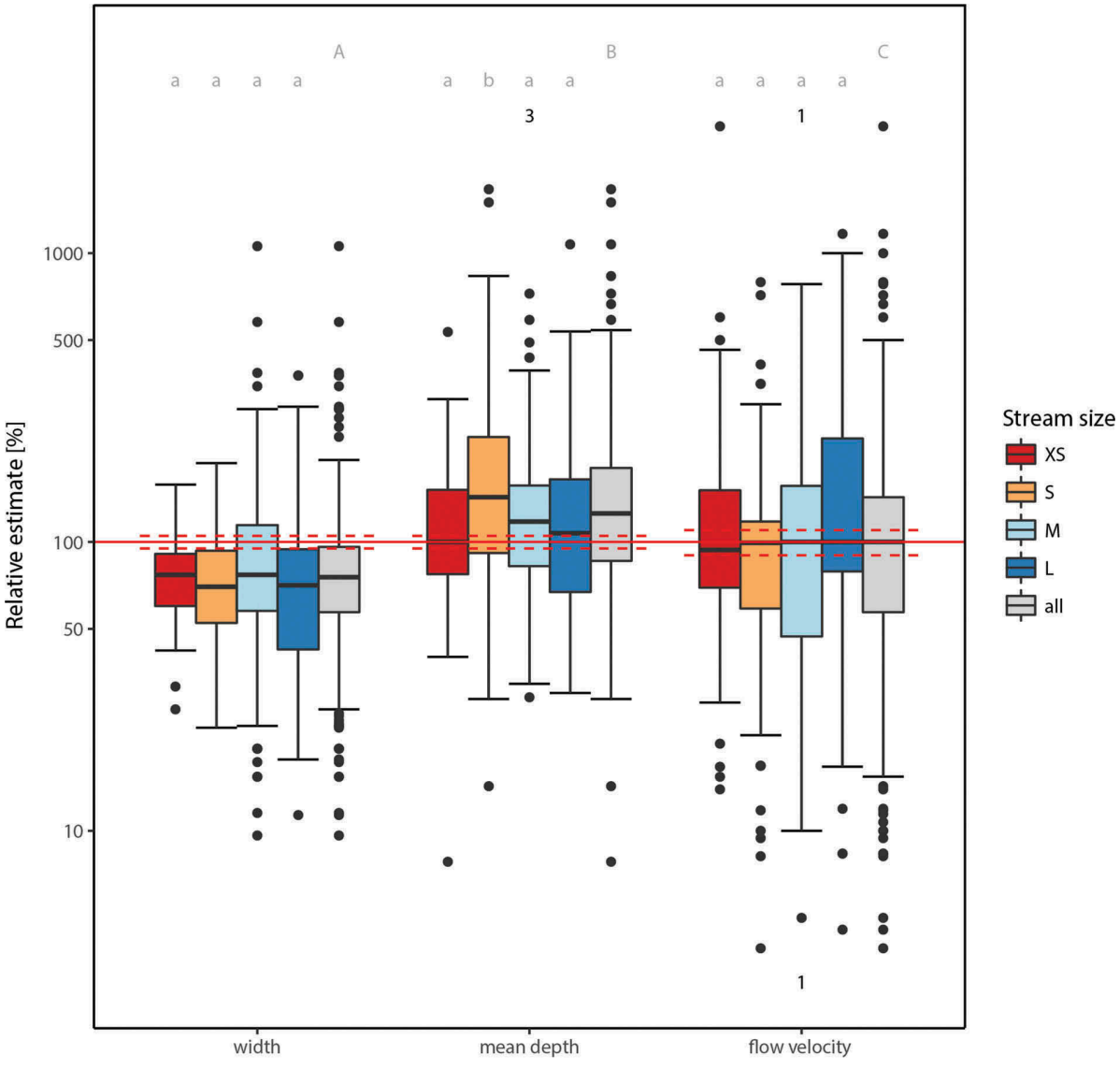


Figure 4. Box plots of the relative estimates of width, mean depth and flow velocity for each stream size class and all streams together. Median relative estimates of width, mean depth and flow velocity of all surveys combined were significantly different (indicated by different upper case letters), whereas between stream size classes they were mostly similar (same lower case letters). The solid red line (100%) indicates that the estimate is the same as the measured value; dashed red lines indicate the 5% (width and mean depth) and 10% (flow velocity) uncertainty bands. The numbers above and below the box plots indicate the number of outliers not shown. Note the log scale.

same pattern, with width having fewer outliers (26%) than flow velocity (39%) and mean depth (41%) (Fig. 4).

3.3 Stream level class estimates

About half of the participants (48%) selected the correct stream level class and most of the remaining participants (40%) were off by only one class. There were only a few outliers (13% of participants had an error of two classes or more; the total does not add to 100% due to rounding) (Fig. 5(a)). The largest overestimation was six classes and the largest underestimation was three classes. These errors likely occurred due to a misunderstanding of the method.

3.4 Comparison of stream level class and streamflow estimates

To allow comparison of the streamflow and stream level class estimates, the latter were translated into corresponding streamflow values. These calculated streamflow values had a narrower interquartile range than the streamflow estimates based on the factors (67–157% compared to 30–163% for Q_{level} and Q_{factor},

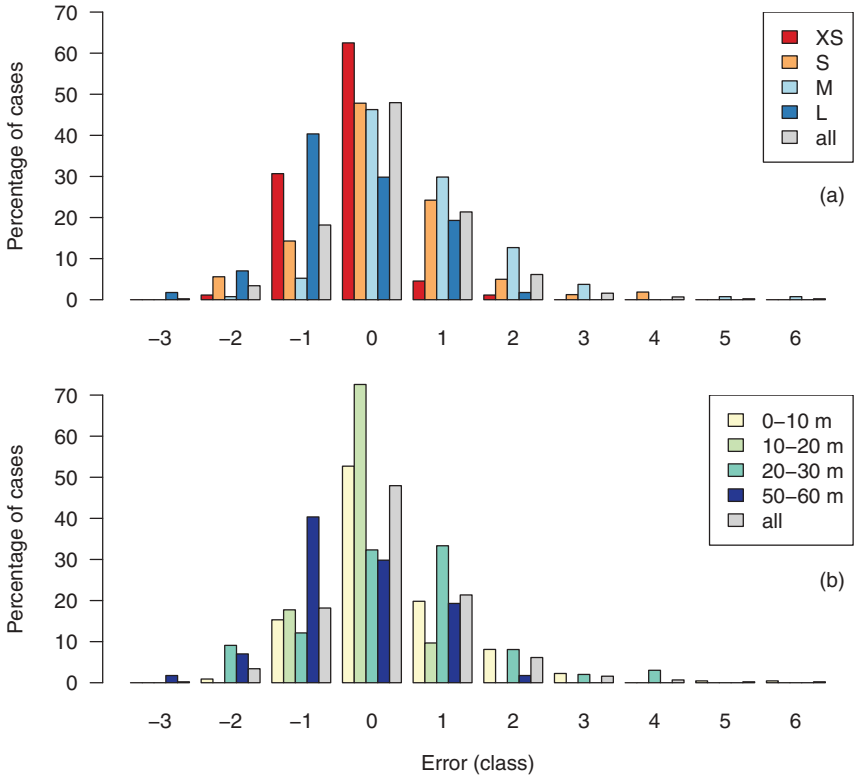


Figure 5. (a) Distribution of errors in stream level class estimates (0: no error, −1: one class lower than the actual stream level class, and 1: one class higher than the actual class) for streams of different sizes; and (b) the distance between participant and the virtual staff gauge, as well as all estimates together. There were no surveys where the virtual staff gauge was 30–50 m away from the participants.

respectively, when all estimates are compared together) and also had fewer outliers (see Fig. 6). Only 39% of the streamflow estimates derived from the stream level class estimates (compared to 66% for Q_{factor}) were significantly overestimated (relative estimate > 150%) or underestimated (relative estimate < 50%). Furthermore, only 3% of the estimates were more than a factor of 10 "off target" (compared to 11% for Q_{factor}). Even when taking the uncertainty in streamflow for the upper and lower stream level class boundaries into account (Fig. 7), the stream level class estimates resulted in streamflow values that were more accurate and had fewer outliers than those determined from the estimated width, mean depth and flow velocity.

Only for the small-sized streams was the interquartile range for streamflow calculated from stream level classes larger than the streamflow determined from the estimated width, depth and flow velocity (Fig. 6). When taking a closer look at the surveys for the different streams, it is clear that mainly the first survey at the Sihl and partly the survey at the Töss caused the large variation in the estimated streamflow from the stream level class data (see Supplementary material, Fig. S3).

3.5 Effect of stream size on streamflow and stream level class estimates

3.5.1 Streamflow

When estimating streamflow directly (Q_{direct}), participants made larger relative errors for the small streams (S; first to third quartile of relative estimates: 55–542%), than for the XS (19–112%), M (23–233%) and L (14–134%) streams. However, general statements on the effect of stream size on the accuracy of streamflow estimates are difficult to make because there were significant differences within each size class as well (Fig. 3).

The interquartile range of the Q_{factor} estimates was significantly smaller for the small (first to third quartile of relative estimates: 49–175%) and medium (27–117%) streams compared to Q_{direct} (Fig. 6). The Q_{factor} estimates were less accurate for XS (interquartile range: 47–293%) and L (17–226%) streams than for S and M streams. For the XS streams this difference is largely based on the estimates from Irchel, where direct streamflow estimates were more accurate than those derived from the estimated factors. For the Hornbach (another XS stream), there was no significant difference between the median relative estimates of Q_{direct} and Q_{factor} (for the Chriesbach

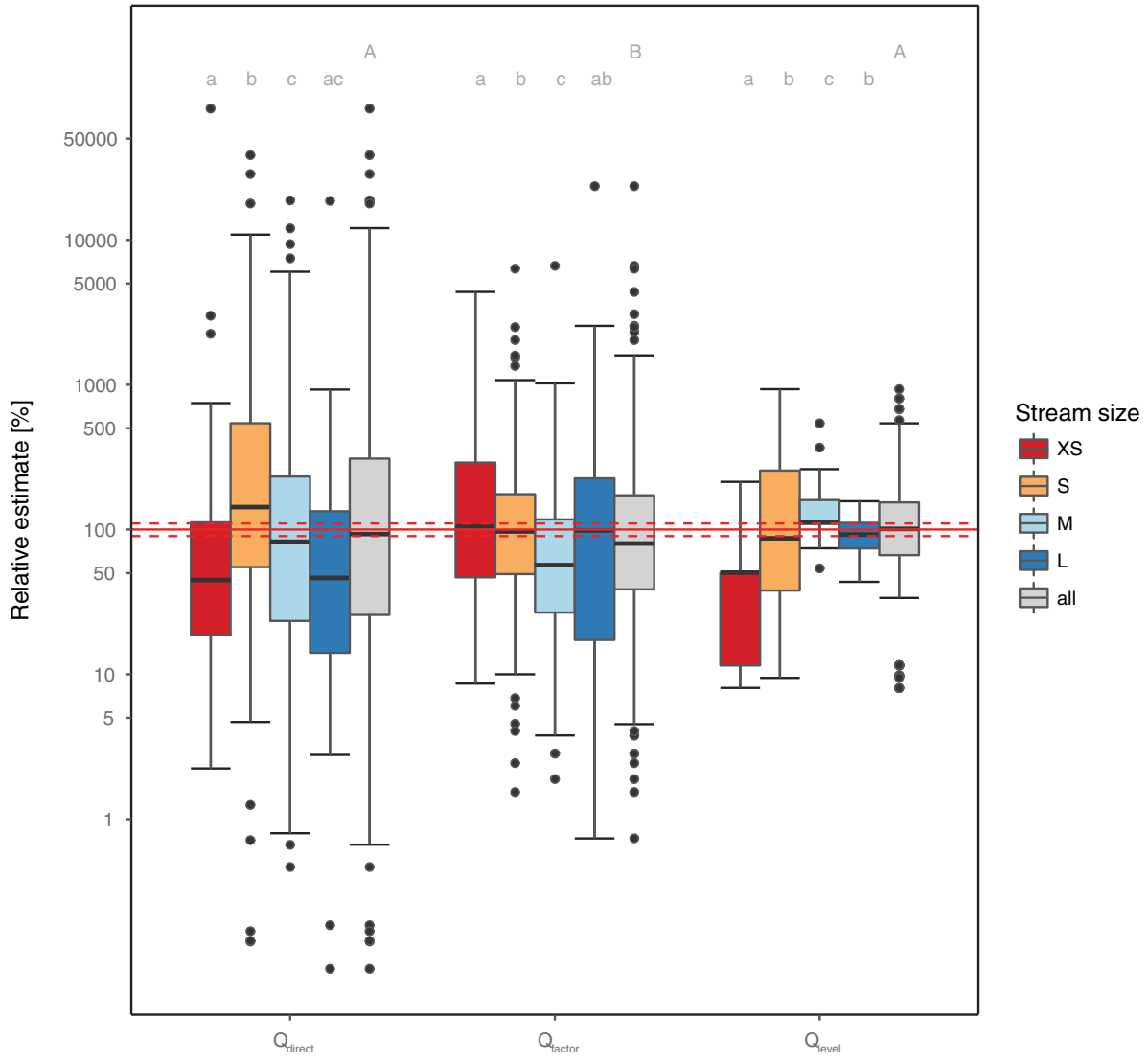


Figure 6. Box plot of the relative estimates of Q_{direct}, Q_{factor} and Q_{level} for each stream size class and all surveys combined. The statistically significant different medians are indicated by different upper case letters (combined data from all surveys) and different lower case letters (per stream size classes). The solid (red) line at 100% indicates that the estimate is the same as the measured value and the dashed (red) lines indicate the 10% uncertainty band for the measured streamflow.

there was no directly estimated streamflow data). The reasons for this different pattern in the Irchel stream are unknown, but could be due to the lower streamflow in the Irchel stream (0.01 m^3/s) compared to the Hornbach (0.13 m^3/s).

3.5.2 Stream level classes

Stream level class estimates were also analysed according to the distance between the participants and the virtual staff gauge, because the distance was not always related to the stream size. For the Limmat the virtual staff gauge was positioned on a bridge pillar rather than the opposite streambank (Fig. 1).

The stream level class estimates were generally more accurate if the staff gauge was closer to the observer (Fig. 5). For a distance of 0–10 m, 53% of participants selected the correct stream level class, while 35% selected a stream level that was only one class away. For a distance of 10–20 m, no one selected a stream level class more than one class from the true value, and 73% of the participants selected the correct class, while for a distance of 20–30 m, 32% of participants were correct and 45% were one class away. For a distance of 50–60 m, 30% of participants chose the correct stream level class and 60% a neighbouring stream level class (Fig. 5(b)). This is not surprising, as, in cases where the

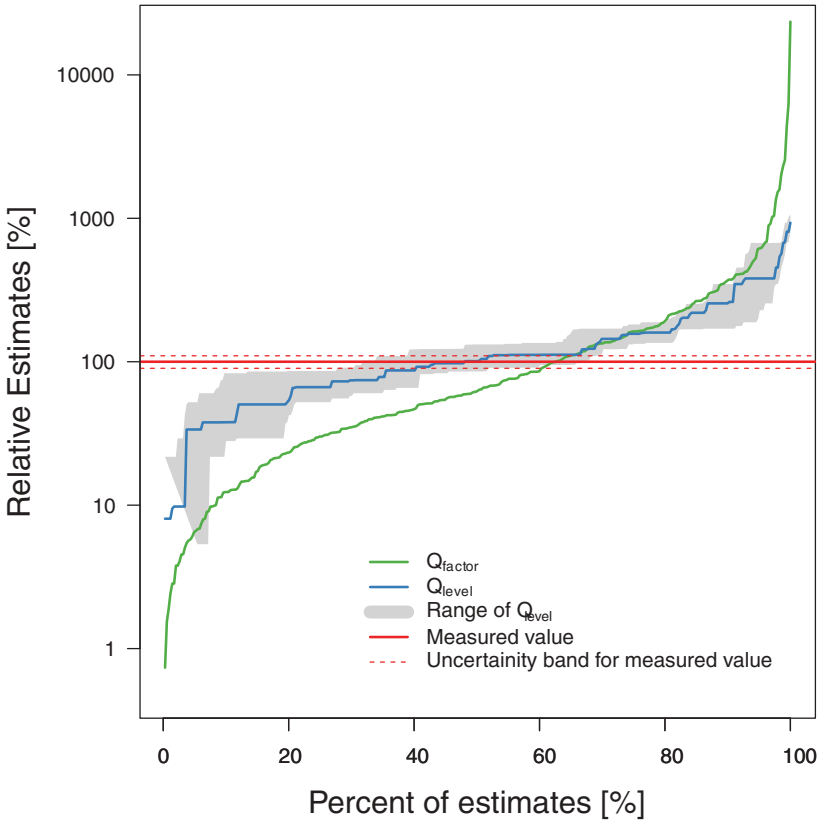


Figure 7. Frequency distribution of the relative streamflow estimates for Q_{factor} and Q_{level}. The shaded (grey) band indicates the upper and lower streamflow for each stream level class. The lower streamflow for each stream level class does not reach the 0% mark, as there were 18 zero values, which cannot be displayed on a log scale.

virtual staff gauge is far away, it is more difficult to discern the stream level class and the reference, such as stones or other helpful objects, on the streambank.

3.6 High *vs* low flow estimates

One issue with hydrological data based on citizen science is the accuracy of the estimated streamflow, but another issue is whether changes in these estimates reflect differences in streamflow over time. Comparison of the estimated streamflow values for the Limmat, Sihl and Aare shows that the median estimated streamflow (Q_{factor}) was higher when the flow was higher, but the differences were not sufficient to fully reflect the increased streamflow (Fig. 8) and were not significant for the Aare (Fig. 8 (b) and (c)). For the Limmat there were significant differences between the surveys, but these differences did not correspond fully to the measured values, as participants underestimated both high and low flow and the differences of estimates between the surveys were seemingly random regardless of high or low flow (Fig. 8(a)).

The variations in streamflow were better represented by the streamflow derived from the stream level class estimates (Q_{level}; Fig. 8(d)–(f)), for which the median estimated streamflow was indeed significantly higher when the flow was higher for seven out of eight surveys. The exception is the median streamflow for the survey on June 2017 at the Limmat, for which the median estimated streamflow (Q_{level}) was not significantly different from the median estimated streamflow during the July and April 2017 surveys, although the first and third quartiles were higher than for the July and April 2017 surveys (see Table 2 and Fig. 8(d)). The variation in streamflow is therefore better represented by streamflow derived from stream level class estimates than by streamflow derived by the factors.

4 Discussion

4.1 Can citizens estimate streamflow accurately?

The results of the streamflow estimation surveys demonstrated the "wisdom of the crowd" effect (Surowiecki 2004, Nielsen 2011) as the median estimates were close to the measured values. However, in practice there will be, at a certain location, only one or at most a few estimates for a certain point in time, so

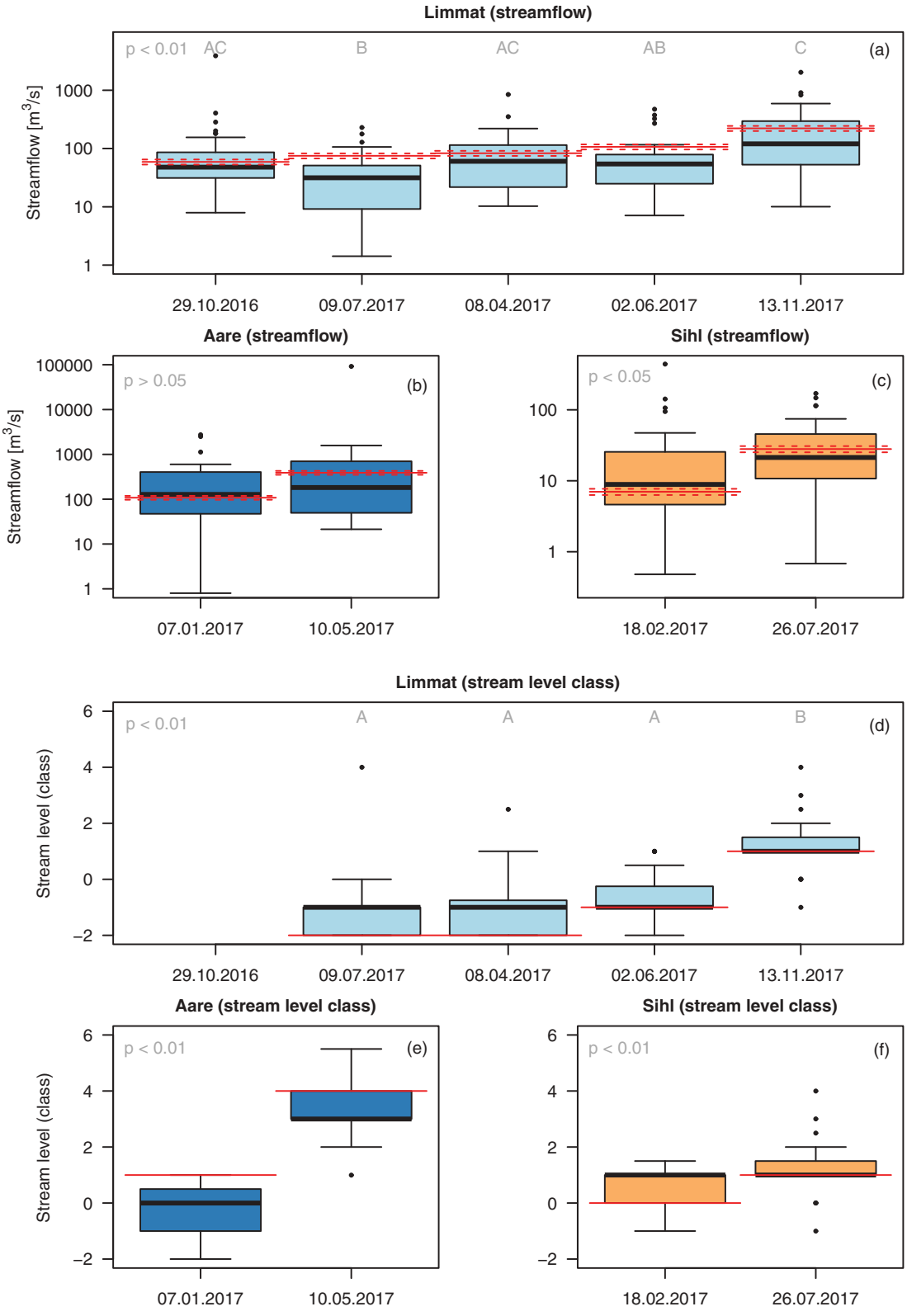


Figure 8. Box plots of (a)–(c) the streamflow based on Q_{factor} and (d)–(f) the estimated stream level classes for different flow conditions for three streams (low flow to high flow in each subplot; see Table 1 for details). Solid and dashed (red) lines as described in Figure 6 caption. The red lines indicate the correct values. Note: the axis ranges are different for each stream. The *p* values indicate the results of the Mann-Whitney (Sihl and Aare) and Kruskal-Wallis (Limmat) tests to determine whether the median estimated streamflow/stream level class of the different surveys are significantly different or not. For the Limmat, surveys with the same upper-case letter (e.g. A) the Dunn *post hoc* test indicated that median streamflow/stream level class estimates were not significantly different from each other.

Table 2. Descriptive statistics of the streamflow derived from the estimated width, mean depth and flow velocity (Q_{factor}; m^3/s) (and relative estimate, %) and the stream level classes for the Aare, Limmat and Sihl for different flow conditions.

Stream	Date	Streamflow, Q_{factor} (m^3/s) (relative Q_{factor}, %)				Stream level class			
		Measured	Percentile			Measured	Percentile		
			25%	50%	75%		25%	50%	75%
Sihl	18.02.2018	7	5	9	26	0	0	1	1
		(100)	(66)	(127)	(365)				
	26.07.2018	28	11	21	46	1	2	2	3
		(100)	(39)	(76)	(163)				
Limmat	29.10.2016	59	31	48	86				
		(100)	(53)	(81)	(146)				
	08.04.2017	83	22	60	111	−2	−2	−1	−1
		(100)	(27)	(73)	(134)				
	02.06.2017	107	26	54	78	−1	−1	−1	0
		(100)	(24)	(51)	72)				
	09.07.2017	75	9	32	49	−2	−2	−1	−1
		(100)	(12)	(42)	(66)				
	13.11.2017	222	53	120	296	1	1	1	2
		(100)	(24)	(54)	(133)				
Aare	07.01.2017	108	47	128	404	0	−1	0	1
		(100)	(44)	(118)	(374)				
	10.05.2017	389	51	182	684	4	3	3	4
		(100)	(13)	(47)	(176)				

for hydrological citizen science projects focusing on streamflow the accuracy of the individual estimates is more important than the accuracy of the median estimate.

As expected, estimation of the individual streamflow factors (width, mean depth and flow velocity) led to more accurate streamflow estimates than the direct estimation of streamflow. The reduction in the number of extreme outliers for estimates based on the streamflow factors is likely due to the more intuitive units in which the estimates have to be given. For non-scientists the unit cubic metres per second (m^3/s) is difficult to visualize and not easy to relate to everyday experiences. Width and depth in metres (m) and flow velocity in metres per second (m/s) are easier to visualize and estimate for most people. The unit litres per second (L/s) is likely more tangible (as one knows the volume of a litre from drink containers and can estimate how long it takes to fill a bottle or a bucket). This might explain why, for the very small Irchel stream, direct streamflow estimates were more accurate than the streamflow derived from the estimated width, depth and velocity, which included the multiplication of three different types of error. For the Hornbach, another very small stream, there was no significant difference between Q_{direct} and Q_{factor}, possibly because it had more streamflow than can fit in a bucket in a second.

The direct streamflow estimates for the Aare (L) were also surprisingly accurate. After the survey, we learned that there used to be a digital display of the current streamflow at the FOEN gauging station, close to the location of our surveys. That display was dismantled before our survey, but it is possible that some participants walked by this site regularly and had a "ballpark" value for the streamflow of the Aare in the back of their minds. Nevertheless, based on our dataset, estimating the streamflow factors rather than the streamflow directly is especially suitable for small and medium streams. It is, however, also important to note that, within the same stream size class, the accuracy of estimates varied for each stream, and even the accuracy of the estimates for the same stream location can vary for different flow conditions (Figs 3 and 8). There was no clear pattern in the relative streamflow estimates (Q_{factor} or Q_{level}) to suggest that either low or high flows are more accurately estimated (see Fig. 8 and Table 2; also supplementary Fig. S4).

Many participants estimated the flow velocity fairly accurately if they threw a twig or leaf into the stream, as we suggested, or even just watched something like a bubble in the stream pass by. The differences between these approaches could not be quantified, as it was not documented who chose which approach.

Even though width and mean depth are measured in the same units, width could be estimated more accurately than mean depth. This is consistent with a study by Wahl (1977), in which trained participants measured both the width and depth of a stream, but measured width with more consistency than depth. In our case this is likely due to the refraction of light in water, as well as the inability to see the bottom of the stream because the water is murky or deep, which was the case for the Sihl at high flow (S), Limmat at high flow (M) and both surveys for the Aare (L). Also in some cases – Hornbach (XS), Irchel (XS), Glatt (S), Sihl (S), Töss (S)

and Limmat (M) – it was feasible to pace the width along a bridge, in order to gain a better estimate, which made the width estimates more accurate; of course this could not be done for depth. According to Gibson and Bergman (1954), distance estimation can be trained and constant over- and underestimation of distances can be improved.

Training is implemented in many citizen science projects to ensure high-quality data (Bonney *et al.* 2009, Haklay *et al.* 2010, See *et al.* 2013, Stepenuck and Genskow 2017). Participants in our survey received no training, had no prior experience and (presumably) only estimated streamflow and its factors once. The effect of a one-time training was tested for some citizen science projects (Crall *et al.* 2013, Rinderer *et al.* 2015) and has been shown to improve the data-collection ability of the participants. Training options for our study could be in the form of online tutorial videos, or a list of well-known streams and their range in streamflow to indicate approximate numbers for streamflow, as well as width, depth and flow velocity. If participants can improve the accuracy of their estimates and the number of outliers can be reduced sufficiently, streamflow estimates might be usable for hydrological model calibration (Etter *et al.* 2018). Further research will test the applicability of quality control methods, such as outlier detection and the effect of training on the accuracy of streamflow estimates.

The inaccuracies of the streamflow estimates should be seen in light of the rating curve errors that are included in conventional measurements, which have a range of ±20% for medium to high flows and substantially higher errors ranging from −60 to +90% for low flows (McMillan *et al.* 2012). Only 29 and 63% of the Q_{direct} estimates were within ±20 and ±90% of the measured streamflow value, respectively. For the Q_{factor} estimates, the respective values were 15 and 73%.

Ensuring, and possibly improving, the accuracy of the crowdsourced data is an important aspect in any citizen science project. The inaccurate estimates of streamflow might be excluded from analyses by quality control methods. A comprehensive overview of data validation methods in the field of citizen science, such as expert review, photo submission or automatic filtering, is provided by Wiggins *et al.* (2011), and many of these methods are likely also applicable to crowdsourced hydrological estimates.

Video imagery is an alternative way to estimate streamflow. These methods have great potential, especially for more accurately determining flow velocities (Bradley *et al.* 2002, Tsubaki *et al.* 2011, Lüthi *et al.* 2014, Le Coz *et al.* 2016, Tauro *et al.* 2018) and have benefits, such as being more objective and possibly allowing a higher accuracy than visual streamflow estimates. By using advanced and sophisticated technology, they also create a curiosity factor that can motivate people. However, there are also some limitations of these approaches in citizen science projects. Issues include light requirements, camera restrictions and the need for initial *in situ* channel measurements as a reference (Lüthi *et al.* 2014). To encourage more participants to join a citizen science project, we were interested to keep the "installation" of new sites and the observation approach as easy as possible. The visual estimates used in this study are easier to apply for many citizens and, thus, can potentially be used to provide more observations. The different methodologies complement each other and different methods might be most suitable for different locations, participant groups or observation goals. Tauro *et al.* (2018) express a similar opinion: "*Reconciling and complementing observations from such an abundant pool of methodologies, devices and platforms is the ultimate goal of the research community towards an improved understanding of hydrological processes*" (Tauro *et al.* 2018, p. 187). Many of the current limitations in video imagery will likely be resolved in the future, making this approach a more usable alternative for streamflow or stream level estimates. A possibility in the future might also be to develop a virtual staff gauge in an augmented reality setting, thereby facilitating participants' stream level class estimates.

4.2 Can citizens estimate stream level classes accurately?

Stream level classes were introduced to simplify the stream level estimation task for the participants. In theory we could have also asked participants to estimate a metric value above or below some fixed point. However, the depth estimates (Fig. 4) for Q_{factor} suggest that this approach would lead to estimates with a low accuracy. The high accuracy of stream level class estimates and the small number of outliers (i.e., estimates that are more than one class off target) indicate that this is a suitable parameter for citizen science projects. The major benefits of the virtual staff gauge approach is that estimates can be done quickly and that relative variations in stream level can be estimated with small uncertainties, but, on the down side, they also have a lower resolution. A participant can be no more than 10 classes off target (which never happened; 0.7% of participants were four classes off and <0.5% of participants were five or six classes off).

Participants only needed to compare the current stream level to a previous stream level using structures,

streambanks or stones as a reference. If the virtual staff gauge is well placed (i.e., there is a suitable structure on the stream bank or in the stream), the participant only needs to look for the reference and then determines the corresponding stream level class. In general, the vast majority of participants had no problem understanding the concept and estimated the stream level class correctly; outliers in the estimated stream level classes were very rare. However, there were also a few clearly wrong stream level class estimates, which might suggest a misunderstanding of the concept by some participants. The two most extreme overestimations were both at the Limmat, the most extreme underestimations at the Aare. Most participants (49%) underestimated the stream level class at the Aare. The reasons are unknown, but potentially this could be attributed to a staff gauge placement during an exceptionally low stream level (less than a 2-year low according to official measurements; BAFU 2017), meaning that the zero value was already very low. This might have confused participants as they may have thought that the staff gauge represents the average streamflow condition.

The stream level class estimates were especially accurate for smaller streams where the opposite stream banks, at which the virtual staff gauges were located in the photo, were close to the participant. The Limmat is a wider stream, but was an exception as the virtual staff gauge was placed on a bridge pillar, which was relatively close to the observer. This is most likely the reason why the stream level class estimates for the Limmat were more accurate than for the Aare (the only stream where the references for the virtual staff gauge were 50–60 m away from the participant), even though the widths of the actual streams were similar (50 and 52 m, respectively). This shows that, for stream level class estimates, the placement of the virtual staff gauge is important. One of the very small streams (Irchel) had a poorly placed staff gauge (the image was taken looking down onto the stream rather than horizontally from the height of the stream level, which distorted the virtual staff gauge relative to the wall behind the stream) and made it more difficult to read. The median relative estimate for Q_{level} for the Irchel stream was 12%, whereas the median relative estimate for Q_{level} for all surveys was 101%.

Several studies have examined the accuracy of crowdsourced data (Haklay *et al.* 2010, Crall *et al.* 2011, See *et al.* 2013, Isaac and Pocock 2015, Tye *et al.* 2016, Aceves-Bueno *et al.* 2017, Mengersen *et al.* 2017), mentioning case studies such as OpenStreetMaps, where Volunteered Geographic Information (VGI) data are collected online and verified by other participants (Haklay *et al.* 2010), and discussing issues such as presence-only data for crowdsourced species classification (Isaac and Pocock 2015, Tye *et al.* 2016, Mengersen *et al.* 2017). While hydrological studies have also discussed crowdsourced data accuracy (Turner and Richter 2011, Rinderer *et al.* 2012, 2015, Lowry and Fienen 2013, Peckenham and Peckenham 2014, Breuer *et al.* 2015, Le Coz *et al.* 2016, Little *et al.* 2016, Weeser *et al.* 2018), most of these studies looked at crowdsourced measurements rather than estimates (Lowry and Fienen 2013, Peckenham and Peckenham 2014, Little *et al.* 2016, Weeser *et al.* 2018). While others, such as Turner and Richter (2011), looked at class estimates, they mainly looked at two class options (wet or dry stream), but unfortunately do not mention data accuracy apart from the fact that participants were trained for consistency. Rinderer *et al.* (2012, 2015), who also looked at classed data, analysed participants' ability to estimate relative soil moisture classes and found that, in one case study, 95% of participants were no more than one class off (Rinderer *et al.* 2012), and in another study with various groups, 81–93% of the participants were no more than one class off (Rinderer *et al.* 2015). However, as far as we are aware, our study is the first to address the accuracy of participants' estimates of stream level classes.

In addition to being more accurate, the stream level class estimation process is also very quick, which is a big advantage for a citizen science project. It is assumed that offering a fast procedure to document stream levels will encourage citizen observers to contribute data to a project regularly (Eveleigh *et al.* 2014). It is very common for citizen science projects that the majority of the contributions come from a small group of high contributors (Lowry and Fienen 2013, Eveleigh *et al.* 2014, Sauermann and Franzoni 2015). For example, in the CrowdHydrology project, one participant walked past a particular station three to four times a week, which led to this station having almost 10 times as many measurements as the station with the next highest number of data submissions (Lowry and Fienen 2013). This highlights the extreme value of these high contributors and shows that it is important to be able to take measurements quickly.

4.3 Are citizens likely to observe variations in streamflow?

Having data for high and low flows, or relative variations in streamflow is crucial in order to determine how a stream reacts to precipitation, snowmelt events or long periods without rainfall, and for hydrological model calibration. Hence, it is important to know if crowdsourced data can properly reflect such variations

in streamflow and whether the accuracy of the data depends on the flow conditions. The results from the surveys suggest that the temporal dynamics in streamflow will be relatively poorly represented by citizen-based streamflow estimates. For two of the three streams (Sihl and Aare), the median streamflow was overestimated at low flows and underestimated at high flows, which indicates insufficient adjustment of the streamflow estimates to the variation in flow conditions. For the Limmat, the significant difference in the streamflow estimates does not seem to correspond to the differences in the measured streamflow (Fig. 8 (a)–(c)). This is partly due to the problem that width (and to a lesser degree velocity) estimates were more accurate compared to depth estimates (Fig. 4). As long as a high flow stays within the streambank, the width of the streams in our survey does not vary significantly between low and high flows. Thus, the majority of the variation in flow conditions is due to the variation in depth, which was most difficult to estimate.

During the surveys we did not ask the same persons to estimate the flow during high and low flow conditions. The results for an individual who reports the streamflow at different times may be different, because the participant might consistently over- or underestimate the flow and therefore the relative variations might be more accurate than indicated by our results (Rinderer *et al.* 2015). Thus, further research is needed to determine if the streamflow dynamics are better described by the streamflow estimates when the majority of the contributions for a particular stream are made by one (or a few) active citizen(s) (Lowry and Fienen 2013).

The high and low flow patterns are better reflected in the stream level class estimates, with the median flow derived from these estimates (Q_{level}) being significantly different between high and low flows for all streams. For the Limmat, the *post hoc* tests showed a significant difference between the high flow and all other survey campaign estimates. This underlines the benefits of collecting stream level class estimates, particularly for model calibration (see additional discussion below).

4.4 Should citizen science projects focus on streamflow or stream level class estimates?

The reduction of the number of outliers in the streamflow estimates calculated from the stream level class data (Q_{level}) compared to the direct streamflow estimates (Q_{direct}) and streamflow estimates based on the streamflow factors (Q_{factor}) can partly be explained by the limited number of potential entries for the virtual staff gauge (i.e., participants can only choose one out of 10 available classes for the stream level estimate). For Q_{direct} and Q_{factor}, participants were able to state any value for their estimates, even values that are physically impossible for a particular stream. Hence, with regard to the reduction of outliers, estimating stream level classes seems advantageous for citizen science projects. Additionally, our results suggest that stream level class estimates appear to be better suited to represent variations in flow conditions. Thus, the results of this study suggest that citizen science projects should focus on stream level class estimates instead of streamflow estimates, although this needs to be tested for different climatic, geographical and socio-economic settings.

However, it should be noted that part of the difference in accuracy for the stream level class estimates and streamflow estimates is due to the difference between relative and absolute values. For our approach, it would be impractical to use classes for streamflow estimates, as we would need many classes, or the resolution of the data would be very low (i.e., the flow for a given stream is likely to always be within the same class). However, as mentioned above, lists of well-known streams, giving their streamflow range to indicate orders of magnitude for the expected streamflow, as well as width, depth and flow velocity, could be provided to make it easier for citizens to make the estimates and to improve the accuracy of the estimates.

One of the disadvantages of the stream level classes is that each class represents a range of potential streamflow values, rather than one specific value. If a participant estimates that the stream level is in class two, it is unclear whether that means the upper, middle or lower part of the class. The other disadvantage is that these estimates do not provide information on streamflow volumes. However, the usability of stream level class data for hydrological model calibration was tested by van Meerveld *et al.* (2017), who showed that stream level class data can be used to calibrate a simple bucket-type hydrological model, and suggested that simple hydrological models can be used to convert stream level class data to time series of streamflow. The value of stream level data for hydrological model calibration, especially for humid catchments, was demonstrated recently by Seibert and Vis (2016). The value of crowdsourced stream level data (photographs of a fixed staff gauge) together with rainfall and flood observations was also shown by Starkey *et al.* (2017). They used community-based observations of rainfall (manual raingauges), river levels (manual staff gauge) and flood-related evidence (anecdotes, photographs or videos) alongside traditional information (tipping bucket raingauge, official raingauge measurements, six

pressure transducers for water level measurements and flow gauging for the discharge-rating curve), in order to fill spatial and temporal gaps in hydrometric data for a 42 km^2 catchment in the UK to improve a physically-based, spatially-distributed catchment model (SHETRAN). Etter *et al.* (2018) calibrated a bucket-type model with synthetic crowdsourced streamflow data with different degrees of error (including errors that are comparable to those observed in this study) and different temporal resolutions, and indeed found that such streamflow estimates do not contain sufficient information to improve the model compared to random parameter sets. However, they also showed that, if the standard deviation of the log-normal distribution that was used to describe the errors of crowdsourced streamflow estimates could be reduced by a factor of two, one estimate per week would lead to a significant improvement in the model simulations.

5 Conclusion

We asked 517 citizens to estimate streamflow directly and indirectly by estimating the stream width, depth and flow velocity. We also asked them to estimate the stream level class. The survey results allowed us to quantify the accuracy of the estimates and are, thus, a basis for evaluating the potential value of citizen science based estimates of streamflow and stream level classes. The median estimated streamflow values were close to the measured streamflow, but there were also many outliers, and the variations in the flow conditions were not fully discernible in the streamflow estimates. The stream level class estimates, which were converted into streamflow values for comparison, had far fewer outliers and were significantly different for the different flow conditions. Stream level class estimates also seemed to be quicker and easier to estimate and are thus considered preferable for citizen science approaches. Hydrological models can then be parameterized based on these stream level class estimates to obtain streamflow time series. The study was conducted in Switzerland and, while we do not expect significant differences, we recommend testing the accuracy of citizen science based estimates of streamflow and stream level classes in different climatic, geographical or socio-economic settings and for rivers with different sizes.

Acknowledgements

We thank all study participants for their time and interest in this research project and for sharing their hydrological estimates with us, as well as the FOEN (Federal Office for the Environment) and WWEA (Office of Waste, Water, Energy and Air of Canton Zurich) for providing the streamflow data used for comparison with the estimates.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was funded by the Swiss National Science Foundation (project 163008, CrowdWater) [Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung 163008, CrowdWater].

ORCID

Barbara Strobl http://orcid.org/0000-0001-5530-4632
Simon Etter http://orcid.org/0000-0002-7553-9102
Ilja van Meerveld http://orcid.org/0000-0002-7547-3270
Jan Seibert http://orcid.org/0000-0002-6314-2124

References

Aceves-Bueno, E., *et al.*, 2017. The accuracy of citizen science data: a quantitative review. *The Bulletin of the Ecological Society of America*, 98 (4), 278–290. doi:10.1002/bes2.1336

BAFU, 2017. *Niedrigwasserwahrscheinlichkeit (Jahresnied rigwasser NM7Q) Aare-Brugg (EDV: 2016)*. Available from: https://www.hydrodaten.admin.ch/lhg/sdi/nq_studien/nq_statistics/2016nq.pdf [Accessed 2 May 2018].

Beven, K. and Westerberg, I., 2011. On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrological Processes*, 25 (10), 1676–1680. doi:10.1002/hyp.v25.10

Beven, K.J., 2012. *Rainfall-runoff modelling: the primer*. 2nd ed. Oxford, UK: Wiley-Blackwell.

Bishop, K., *et al.*, 2008. Aqua Incognita: the unknown headwaters. *Hydrological Processes*, 22, 1239–1242. doi:10.1002/hyp.7049

Bonney, R., *et al.*, 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, 59 (11), 977–984. doi:10.1525/bio.2009.59.11.9

Bradley, A.A., *et al.*, 2002. Flow measurement in streams using video imagery. *Water Resources Research*, 38 (12). doi:10.1029/2002WR001317

Breuer, L., *et al.*, 2015. HydroCrowd: a citizen science snapshot to assess the spatial control of nitrogen solutes in surface waters. *Scientific Reports*, 5, 16503. doi:10.1038/srep16503

Buytaert, W., *et al.*, 2014. Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, 2 (26), 21. Available from: http://journal.frontiersin.org/article/10.3389/feart.2014.00026/abstract

Crall, A.W., *et al*., 2011. Assessing citizen science data quality: An invasive species case study. *Conservation Letters*, 4 (6), 433–442. doi:10.1111/j.1755-263X.2011.00196.x

Crall, A.W., *et al*., 2013. The impacts of an invasive species citizen science training program on participant attitudes, behavior, and science literacy. *Public Understanding of Science*, 22 (6), 745–764. doi:10.1177/0963662511434894

Dickinson, J.L., Zuckerberg, B., and Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics*, 41 (1), 149–172. doi:10.1146/annurev-ecolsys-102209-144636

Dingman, S.L., 2015. *Physical Hydrology*. 3rd ed. Long Grove: Waveland Press Inc.

Dunn, O.J., 1964. Multiple Comparisons Using Rank Sums. *Technometrics*, 6 (3), 241–252. doi:10.1080/00401706.1964.10490181

Engel, S.R. and Voshell, J.R., 2002. Volunteer biological monitoring: can it accurately assess the ecological condition of streams? *American Entomologist*, 48, 164–177. doi:10.1093/ae/48.3.164

Etter, S., *et al*., 2018. Value of uncertain streamflow observations for hydrological modelling. *Hydrology and Earth System Sciences*, 22, 5243–5257. doi:10.5194/hess-22-5243-2018

Eveleigh, A., *et al*., 2014. Designing for dabblers and deterring drop-outs in citizen science. *In*: *CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 26 April–1 May, Toronto. New York, NY: Association for Computing Machinery, 2985–2994. doi:10.1145/2556288.2557262

Field, A., Miles, J., and Field, Z., 2013. *Discovering statistics using R*. Los Angeles: Sage.

Gibson, E.J. and Bergman, R., 1954. The effect of training on absolute estimation of distance over the ground. *Journal of Experimental Psychology*, 48 (6), 473–482. doi:10.1037/h0055007

Hadj-Hammou, J., *et al*., 2017. Getting the full picture: Assessing the complementarity of citizen science and agency monitoring data. *PLoS ONE*, 12 (12), 1–18. doi:10.1371/journal.pone.0188507

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37 (4), 682–703. doi:10.1068/b35097

Haklay, M., (Muki), *et al*., 2010. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47 (4), 315–322. doi:10.1179/000870410X12911304958827

Herschy, R.W., 1971. *The magnitude of errors at flow measurement stations*. Technical report, Water Resources Board, Reading, UK.

Isaac, N.J.B. and Pocock, M.J.O., 2015. Bias and information in biological records. *Biological Journal of the Linnean Society*, 115 (3), 522–531. doi:10.1111/bij.12532

Juston, J., Seibert, J., and Johansson, P., 2009. Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. *Hydrological Processes*, 23 (21), 3093–3109. doi:10.1002/hyp.7421

Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, 41, W03S04. doi:10.1029/2005WR004362

Kundzewicz, Z.W., 1997. Water resources for sustainable development. *Hydrological Sciences Journal*, 42 (4), 467–480. doi:10.1080/02626669709492047

Le Coz, J., *et al*., 2016. Crowdsourced data for flood hydrology: feedback from recent citizen science projects in Argentina, France and New Zealand. *Journal of Hydrology*, 541, 766–777. doi:10.1016/j.jhydrol.2016.07.036

Little, K.E., Hayashi, M., and Liang, S., 2016. Community-Based Groundwater Monitoring Network Using a Citizen-Science Approach. *Groundwater*, 54 (3), 317–324. doi:10.1111/gwat.2016.54.issue-3

Lowry, C.S. and Fienen, M.N., 2013. CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists. *Ground Water*, 51 (1), 151–156. doi:10.1111/j.1745-6584.2012.00956.x

Lüthi, B., Philippe, T., and Peña-Haro, S., 2014. Mobile device app for small open-channel flow measurement. *In*: D.P. Ames, N.W.T. Quinn, and A.E. Rizzoli, eds. *7th International Congress on Environmental Modelling and Software*. San Diego, CA. Available from: http://www.iemss.org/sites/iemss2014/papers/iemss2014_submission_112.pdf

Manning, R., 1891. On the flow of water in open channels and pipes. *Transactions of the Institution of Civil Engineers of Ireland*, 20, 161–207.

McMillan, H., Krueger, T., and Freer, J., 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*, 26 (26), 4078–4111. doi:10.1002/hyp.v26.26

Mengersen, K., *et al*., 2017. Modelling imperfect presence data obtained by citizen science. *Environmetrics*, 28 (5), 1–29. doi:10.1002/env.2446

Nielsen, M., 2011. *Reinventing Discovery: The New Era of Networked Science*. Princeton, NJ: Princeton University Press.

Peckenham, J.M. and Peckenham, S.K., 2014. Assessment of quality for middle level and high school student-generated water quality data. *Journal of the American Water Resources Association*, 50 (6), 1477–1487. doi:10.1111/jawr.12213

Pelletier, P.M., 1988. Uncertainties in the single determination of river discharge: a literature review. *Canadian Journal of Civil Engineering*, 15 (5), 834–850. Available from: http://www.nrcresearchpress.com/doi/10.1139/l88-109

Perrin, C., *et al*., 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. *Hydrological Sciences Journal*, 52 (1), 131–151. Available from: http://www.tandfonline.com/doi/abs/10.1623/hysj.52.1.131

Pool, S., Viviroli, D., and Seibert, J., 2017. Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, 554, 613–622. doi:10.1016/j.jhydrol.2017.09.037

Rinderer, M., *et al*., 2012. Sensing with boots and trousers - qualitative field observations of shallow soil moisture

patterns. *Hydrological Processes*, 26 (26), 4112–4120. Available from: 10.1002/hyp.9531 [Accessed 27 Mar 2014].

Rinderer, M., *et al*., 2015. Qualitative soil moisture assessment in semi-arid Africa - the role of experience and training on inter-rater reliability. *Hydrology and Earth System Sciences*, 19, 3505–3516. doi:10.5194/hess-19-3505-2015

Ruhi, A., Messager, M.L., and Olden, J.D., 2018. Tracking the pulse of the Earth's fresh waters. *Nature Sustainability*, 1 (4), 198–203. doi:10.1038/s41893-018-0047-7

Sauermann, H. and Franzoni, C., 2015. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, 112 (3), 679–684. doi:10.1073/pnas.1408907112

See, L., *et al*., 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PlosONE*, 8 (7), 1–11. doi:10.1371/journal.pone.0069958

Seibert, J. and Beven, K.J., 2009. Gauging the ungauged basin: how many discharge measurements are needed? *Hydrology and Earth System Sciences*, 13 (6), 883–892. Available from: http://www.hydrol-earth-syst-sci.net/13/883/2009/

Seibert, J. and McDonnell, J.J., 2015. Gauging the ungauged basin: relative value of soft and hard data. *Journal of Hydrologic Engineering*, 20 (1), A4014004-1–6. Available from: https://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000861

Seibert, J. and Vis, M.J.P., 2016. How informative are stream level observations in different geographic regions? *Hydrological Processes*, 30 (14), 2498–2508. doi:10.1002/hyp.10887

Starkey, E., *et al*., 2017. Demonstrating the value of community-based ('citizen science') observations for catchment modelling and characterisation. *Journal of Hydrology*, 548, 801–817. doi:10.1016/j.jhydrol.2017.03.019

Statistik Stadt Zürich, 2017. Statistisches Jahrbuch der Stadt Zürich 2017, 188–201. Available from: https://www.stadt-zuerich.ch/prd/de/index/statistik/publikationen-angebote/publikationen/Jahrbuch/statistisches-jahrbuch-der-stadt-zuerich_2017.html [Accessed 12 Oct 2017].

Stepenuck, K.F. and Genskow, K.D., 2017. Characterizing the breadth and depth of volunteer water monitoring programs in the united states. *Environmental Management*, 61 (1), 46–57. doi:10.1007/s00267-017-0956-7

Surowiecki, J., 2004. *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. London, UK: Little Brown.

Tauro, F., *et al*., 2018. Measurements and observations in the XXI century (MOXXI): Innovation and multi-disciplinarity to sense the hydrological cycle. *Hydrological Sciences Journal*, 63 (2), 169–196. doi:10.1080/02626667.2017.1420191

Tsubaki, R., Fujita, I., and Tsutsumi, S., 2011. Measurement of the flood discharge of a small-sized river using an existing digital video recording system. *Journal of Hydro-Environment Research*, 5 (4), 313–321. doi:10.1016/j.jher.2010.12.004

Tulloch, A.I.T., *et al*., 2013. Realising the full potential of citizen science monitoring programs. *Biological Conservation*, 165, 128–138. doi:10.1016/j.biocon.2013.05.025

Turner, D.S. and Richter, H.E., 2011. Wet/dry mapping: using citizen scientists to monitor the extent of perennial surface flow in dryland regions. *Environmental Management*, 47 (3), 497–505. doi:10.1007/s00267-010-9607-y

Tye, C.A., *et al*., 2016. Evaluating citizen versus professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology*, 54 (2), 628–637.

van Meerveld, H.J., Vis, M.J.P., and Seibert, J., 2017. Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences*, 21 (9), 4895–4905. doi:10.5194/hess-21-4895-2017

Vis, M., *et al*., 2015. Model calibration criteria for estimating ecological flow characteristics. *Water (Switzerland)*, 7 (5), 2358–2381.

Wahl, K.L., 1977. Accuarcy of channel measurements and the implications in estimating streamflow characteristics. *Journal Research of the U.S. Geological Survey*, 5 (6), 811–814.

Weeser, B., *et al*., 2018. Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. *Science of The Total Environment*, 632, 1590–1599. doi:10.1016/j.scitotenv.2018.03.130

Welber, M., *et al*., 2016. Field assessment of noncontact stream gauging using portable surface velocity radats (SVR). *Water Resources Research*, 52, 1108–1126. doi:10.1002/2015WR017906

Westerberg, I., *et al*., 2011. Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 25 (4), 603–613. doi:10.1002/hyp.7848

Wiggins, A., *et al*., 2011. Mechanisms for data quality and validation in citizen science. *In*: *Seventh IEEE International Conference on e-Science Workshops*, 5–8 December. Stockholm: IEEE, 14–19. doi:10.1109/eScienceW.2011.27

PAPER III

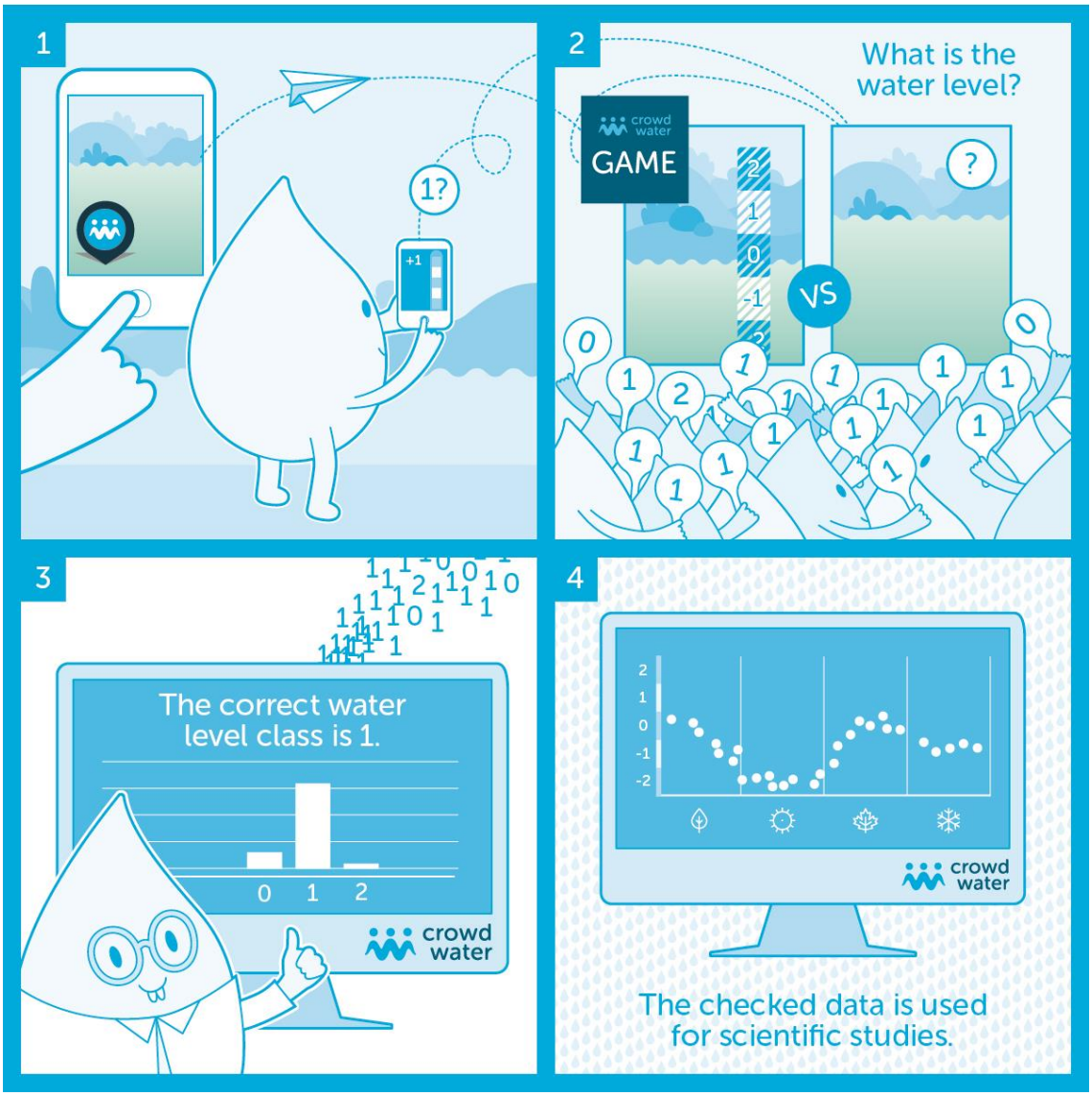


Figure by Tara von Grebel

Strobl, B., S. Etter, H.J. van Meerveld, and J. Seibert (2019), The CrowdWater Game: a playful way to improve the accuracy of crowdsourced water level class data, *PLoS One*, https://doi.org/10.1371/journal.pone.0222579.

PLOS ONE

RESEARCH ARTICLE

The CrowdWater game: A playful way to improve the accuracy of crowdsourced water level class data

Barbara Strobl[1]*, **Simon Etter**[1], **Ilja van Meerveld**[1], **Jan Seibert**[1,2]

1 Department of Geography, University of Zurich, Zurich, Switzerland, **2** Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

* barbara.strobl@geo.uzh.ch

OPEN ACCESS

Citation: Strobl B, Etter S, van Meerveld I, Seibert J (2019) The CrowdWater game: A playful way to improve the accuracy of crowdsourced water level class data. PLoS ONE 14(9): e0222579. https://doi.org/10.1371/journal.pone.0222579

Editor: Seyedali Mirjalili, Torrens University Australia, AUSTRALIA

Received: April 5, 2019

Accepted: September 2, 2019

Published: September 26, 2019

Data Availability Statement: All data files are available from the Zenodo data repository (https://doi.org/10.5281/zenodo.2630587).

Funding: This study was funded by the Swiss National Science Foundation (www.snf.ch; project 163008, CrowdWater). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Data quality control is important for any data collection program, especially in citizen science projects, where it is more likely that errors occur due to the human factor. Ideally, data quality control in citizen science projects is also crowdsourced so that it can handle large amounts of data. Here we present the CrowdWater game as a gamified method to check crowdsourced water level class data that are submitted by citizen scientists through the CrowdWater app. The app uses a virtual staff gauge approach, which means that a digital scale is added to the first picture taken at a site and this scale is used for water level class observations at different times. In the game, participants classify water levels based on the comparison of the new picture with the picture containing the virtual staff gauge. By March 2019, 153 people had played the CrowdWater game and 841 pictures were classified. The average water level for the game votes for the classified pictures was compared to the water level class submitted through the app to determine whether the game can improve the quality of the data submitted through the app. For about 70% of the classified pictures, the water level class was the same for the CrowdWater app and game. For a quarter of the classified pictures, there was disagreement between the value submitted through the app and the average game vote. Expert judgement suggests that for three quarters of these cases, the game based average value was correct. The initial results indicate that the CrowdWater game helps to identify erroneous water level class observations from the CrowdWater app and provides a useful approach for crowdsourced data quality control. This study thus demonstrates the potential of gamified approaches for data quality control in citizen science projects.

1. Introduction

Data quality and quality control are frequently discussed for citizen science projects because these data are generally perceived to be less accurate than traditional data due to human errors. Nonetheless several studies have shown that citizen science data can be as accurate as data from experts [1–3]. Data quality control in citizen science has several purposes, the most

obvious being the improvement of the data quality. Additionally, improved data quality also increases the credibility of the data for the users and the confidence of the citizen scientists in their ability to submit useful data [4,5].

Different data quality control approaches have been developed for citizen science projects within different scientific fields and for different data collection approaches [4,6–8]. Wiggins et al. [8] summarised 18 approaches for data quality control, which can be grouped into approaches before, during and after data collection. These include training participants and providing tutorial materials [9,10], filtering of the incoming data based on the plausibility of the data and the likelihood for a particular geographic region [6,10–14], bias correction, for example for presence only data [10,11,15,16], and review of incoming data [4,8,11,17,18].

The review approach includes reviews by professional scientists, reviews by experienced contributors or regional experts, and peer-reviews by multiple parties [8]. The three review approaches can also be combined within a project, e.g. by asking the public to flag certain entries (peer-review), which are then reviewed by experts [6,18]. Review through professionals is a time-consuming task and can only be done in citizen science projects with a limited amount of incoming data [4]. While all non-fully automated data quality control approaches might be time-consuming, the effort becomes more doable if it can be shared by many people. Review by experienced contributors or regional experts can therefore also be used for large projects. However, it opens the question whom to assign this *"ambassador-status"*, and depending on the data volume and number of ambassadors it might still be a lot of work for a few dedicated volunteers [11,18]. Peer-review by multiple parties is a method to crowdsource data quality control. The quality of the review is insured through multiple assessments so that individual mistakes or misclassifications are insignificant when all assessments are taken into account. Through peer review, the quality control mechanism is scalable, so that it can even be used for big projects. Furthermore, it ensures that citizen scientists are involved in both data collection and data quality control.

There are many examples of peer-review in the field of citizen science, such as projects related to Volunteered Geographic Information [7,19], such as OpenStreetMaps [20], in projects where volunteers make visual comparisons of spatial patterns, such as Pattern Perception [21], Cyclone center [22] and Galaxy Zoo [23], and in projects where volunteers assess pictures, such as Snapshot Serengeti [24] and Cropland Capture [17]. As with many citizen science projects and tasks, a major difficulty associated with the peer-review approach is the recruitment and retention of a sufficient number of reviewers [17,21–25]. Depending on the project, the scientific field and the specific task at hand, different strategies can be employed. A frequently applied strategy, especially for online citizen science projects, is the gamification of tasks [26–28].

Gamification can range from simple points and leaderboards to more immersive games with complex storylines [28]. Different phrases are used in the literature for these types of games, such as citizen science games [29,30], knowledge games [31], games with a purpose [32,33] or serious games [29]. Examples of projects that gamified their interaction with citizen scientists are Foldit, StallCatchers, Phylo, Serengeti Pictures and Cropland Capture. A comprehensive list of gamified citizen science projects can be found on www.citizensciencegames.com.

Many of these games were successful in finding a large number of participants: $>$ 2 500 players in Cropland Capture [34], $>$ 12 000 players in Phylo [35], and $>$ 57 000 players in Foldit [36]. The different topics of the games make comparisons between them difficult, but most publications describe the games as a success. The project Foldit mentions that their players can *"produce structure solutions of the highest quality"* [37] and Curtis [38] says that *"the games [Foldit, Phylo and EteRNA] have the potential to greatly improve our understanding of the*

genetic processes underlying important diseases". For Cropland Capture, the conclusions were slightly mixed *"At first glance [. . .] volunteers are highly effective at rating photographs and satellite imagery for the presence of cropland."*, but *"extracting a reliable signal from crowdsourced data without guidance from expert validations is not possible for this type of task."* [34].

This paper focuses on the value of the online CrowdWater game (https://crowdwater.ch/en/crowdwater-game/) to check and improve the accuracy of crowdsourced water level class data submitted via the CrowdWater app. The CrowdWater game can be described as a citizen science game for which the primary purpose is data quality control rather than education. One of the main differences between the CrowdWater game and most other citizen science games is the complementarity of the tasks (collection and correction) in the CrowdWater project. The CrowdWater project asks citizens to submit water level class data for streams and rivers through an app [39], which in turn are checked by (other) participants through the online CrowdWater game. Therefore, unlike games such as Foldit, Phylo or Cropland Capture, the CrowdWater game does not produce data, but checks the quality of the crowdsourced data. This means that there are two potential entry points into the project (the app and the game; Fig 1) and that there is a range of tasks and interactions available for participants. This might help to *"sustain engagement over time"* [27]. iNaturalist [[40] and iSpot [41] have a similar complementarity as the CrowdWater project, by asking the citizen scientists to collect pictures of plants and animals and by also asking citizen scientists with good species recognition abilities to help classify these pictures.

The specific research questions for this study were:

1. Can the CrowdWater game be used to correct mistakes in the data submitted through the CrowdWater app?
2. Can players correctly identify unsuitable observations through the report function in the game?
3. Is the assignment of the water level class by regular players more accurate than for novice players?
4. What motivates participants to play the CrowdWater game?

2. Methods

2.1 The CrowdWater app

The CrowdWater app [39,42] can be used by citizen scientists worldwide to collect hydrological data. Currently, the app can be used to collect data for three parameters: water level class (and as an advanced option, streamflow), soil moisture, and the occurrence of flow in temporary streams. The CrowdWater game was developed as a data quality control mechanism for the water level class data. To report changes in water levels with the app, users first have to create a reference picture, which is a picture of a stream with a virtual staff gauge that is inserted digitally onto the picture like a sticker. Hence the staff gauge only exists in the reference picture and no physical installations are needed. The size of the virtual staff gauge is controlled by the user but the number of classes is fixed at ten. The virtual staff gauge is placed in the picture in such a way that per definition the water level in the reference picture is always at class zero. At a later time, the user who has made the reference picture or any other citizen scientist visiting the same location can look at the initial reference picture and compare the water level in the current situation with that on the reference picture. They select the water level class that they think best represents the current situation and upload a picture of the new situation. This is called *"observation"* and results in a time series of water level class data with relative values of

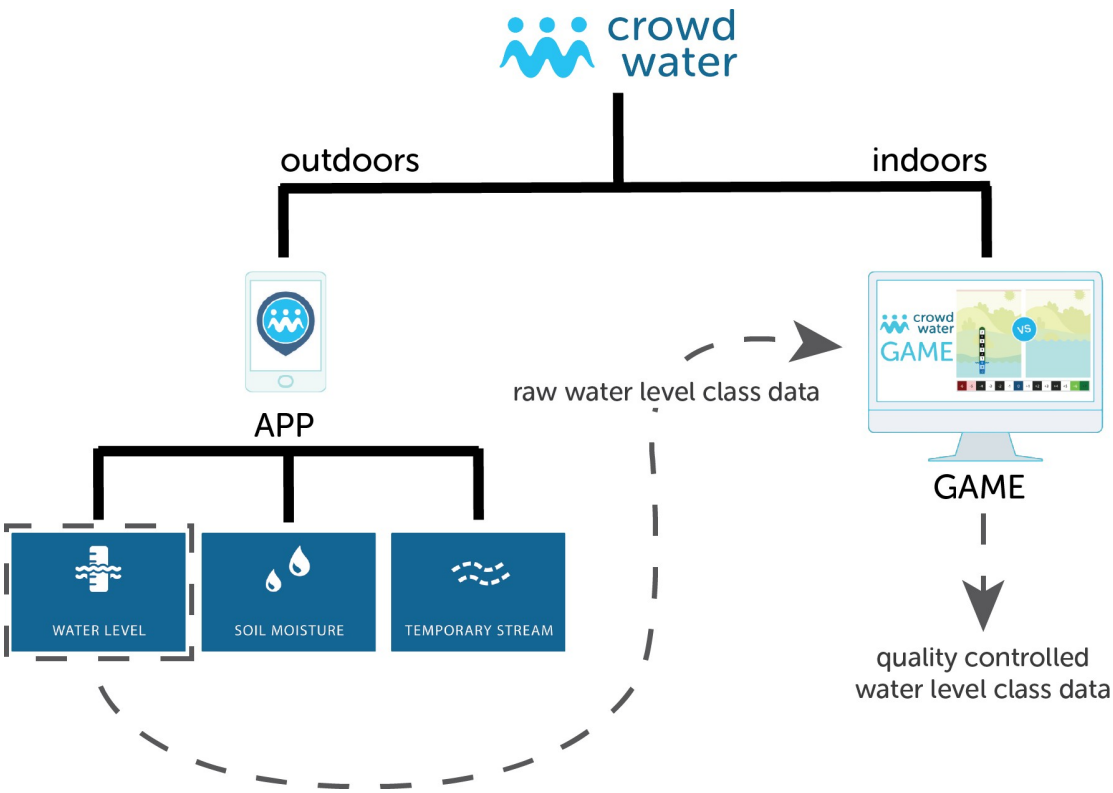


Fig 1. Schematic overview of the CrowdWater project, showing the connection between the CrowdWater app and game.

https://doi.org/10.1371/journal.pone.0222579.g001

the water level for that particular location, rather than time series of actual (metric) water level values. For further information on the CrowdWater app, as well as the use of the virtual staff gauge approach, we refer the reader to the following publications [39,43–45].

2.2 The CrowdWater game

2.2.1 The game. The CrowdWater game is a casual game, which means that very little time, background knowledge, experience or training is needed to start playing the game [46]. The instruction manual is available on the homepage of the CrowdWater project and can be read easily in 5 to 10 minutes. In addition, there is a short (<2 min) movie explaining the game (https://crowdwater.ch/en/instructions/). The task for the players can be described as a series of "*microtasks*", which refers to "*[. . .] systems [that] achieve high quality, typically as good as or better than expert annotators, through extensive use of redundancy and aggregation.*" [47]. The gamification aspects of the CrowdWater game include championships, rounds, points and leaderboards.

The crowdsourced water level class observations displayed in the CrowdWater game are obtained from the CrowdWater app (see Fig 1 and 2.1 The CrowdWater app). The CrowdWater game uses all water level class observations with a picture and displays each picture together with the reference picture for that site (Fig 2). By February 28, 2019, there were 2326 picture pairs in the game but this number is increasing continuously as the app users continue to create new spots with reference pictures and provide observations and pictures for existing spots. These pictures are automatically transferred to the game. The game players compare the picture pairs and estimate the water level class for the picture of the new observation (i.e. the

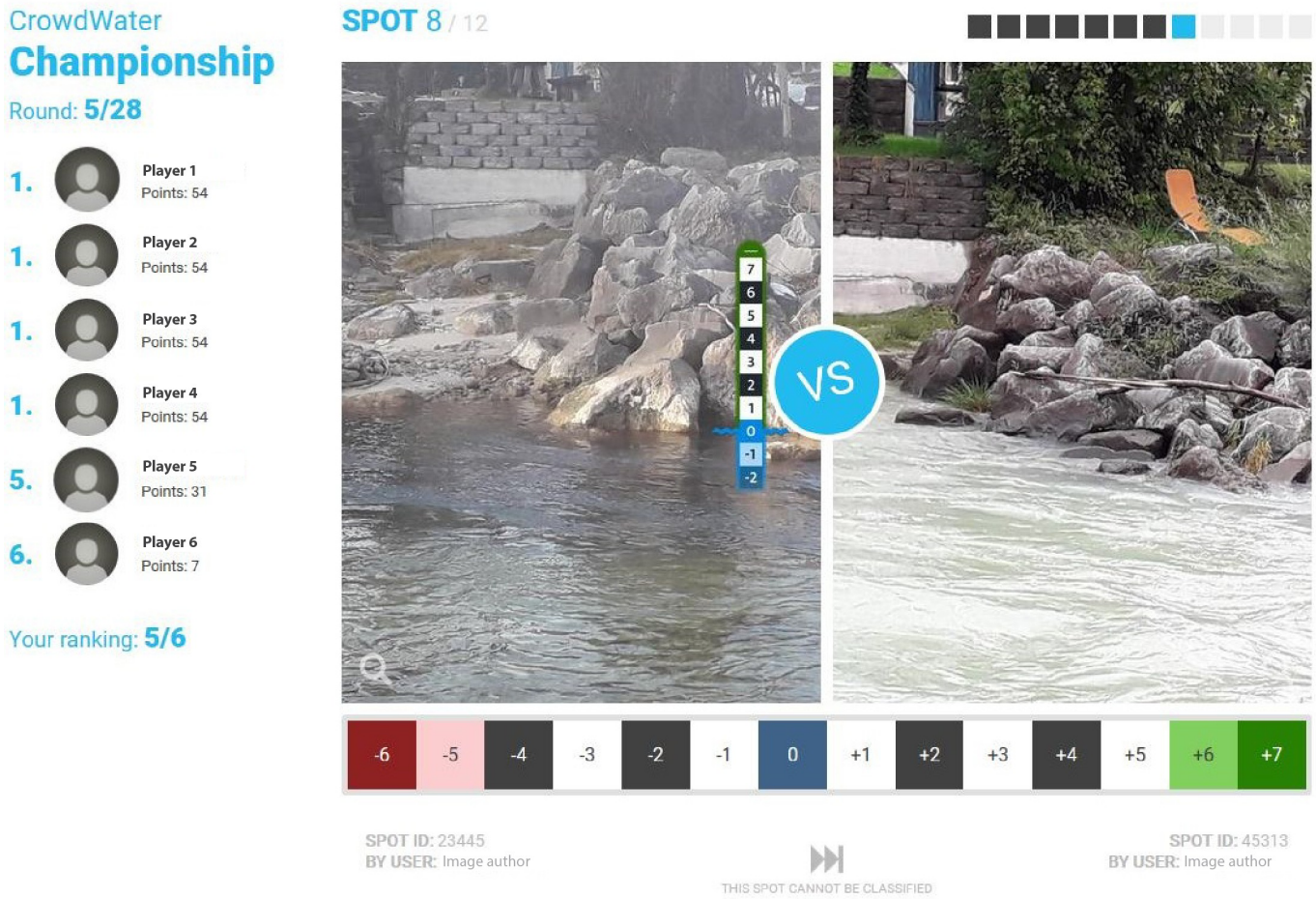


Fig 2. Screenshot of the CrowdWater game. The left picture shows the reference picture with the virtual staff gauge. The right picture shows an observation for the same spot at a different time. The player has to estimate the water level class for the picture on the right by comparing the water level and features in the stream or on the stream bank for both pictures. On the left hand side the scores of the top ten players for this round of the game are shown. The report button at the bottom can be used for pictures that cannot be classified. The squares at the top indicate the number of comparisons completed in this round of the game so far (black) and comparisons still to come (grey).

https://doi.org/10.1371/journal.pone.0222579.g002

one that does not have the virtual staff gauge). This way many players can assess the same situation without being outside at the same stream at the same time and thereby peer-review the water level class data that are submitted via the app. As the players only get to see a photograph, rather than seeing the actual stream, we need to test the value of the game for data quality control and how many votes need to be collected per observation. Obviously, if a water level observation is uploaded without a new picture (< 3% of all observations), data quality control via the game is not possible.

2.2.2 Observation reports. While playing the game, players can use a report button (Fig 2) when they believe that the water level cannot be determined for the new picture or when there are other issues with the pictures. Players can choose one of six reasons to report an observation. In the case of *"other reason"* the player can write what that reason is.

- The photo is ok, but I don't know the category.
- The staff gauge is not placed correctly.

- The staff gauge is missing.
- The approved value is clearly incorrect.
- The location has changed and the reference image is unrecognizable.
- Other reason: . . .

2.2.3 Gamification aspects of the CrowdWater game. In its current implementation, the CrowdWater game has monthly championships that consist of 28 daily rounds. Twelve picture pairs are shown to the players per round. There are two different types of picture pairs: classified and unclassified observations. Classified observations have already received 15 or more votes by at least 15 different players. On February 28, 2019, there were 846 classified observations (i.e., classified pictures) in the game. Within the game, the median of the votes for the water level class for a classified observation is calculated and used as the approved value to which the vote of the current player is compared. Currently, the classified observations remain in the game until they have received 100 votes. When there are fewer than 15 votes for a picture, it is defined as *"unclassified"* and, thus, does not have an approved value assigned to it yet. On February 28, 2019, there were 1480 unclassified observations in the game. The players do not know which type of observation they are looking at until they have voted for a water level class. This is a similar approach as for the Cropland Capture game [17]. For already classified observations, players receive six points if they choose the same class as the median of the votes from the other players (considered to be the approved class), four points if they choose a neighbouring class and zero points if they are more than one class off from the median. For (so far) unclassified observations, players always receive three points (regardless of their vote). When reporting a problem for a picture pair (see section 2.2.2 Observation reports) players always receive three points so that there is no incentive to try to classify a picture that should be reported. The distribution of points also ensures that it is not possible to win a round merely by reporting every observation.

After completing a full round of pictures in the game, the players see the score and all twelve picture pairs together with his/her vote and the approved water level class for each picture pair. This provides feedback to the player on what the correct water level class was. At this stage it is possible to report a picture again (but this time this does not lead to any points).

The names of the top ten players for the daily rounds are shown on the leader board (Fig 2). Every month a new championship starts and small prizes are given to the overall winners of the monthly championship (i.e. the three players with the most points for that month), as well as three randomly selected players who won at least one of the daily-rounds during the championship.

2.2.4 CrowdWater game participants. The game was tested internally (i.e., within our research group) between May and July 2018 and has been published and promoted online since August 2018. The game was advertised through several communication channels: the CrowdWater homepage, facebook, twitter, LinkedIn, ResearchGate, CrowdWater app push-messages, CrowdWater newsletter, SciStarter, Schweiz Forscht, citizensciencegames.com, as well as by directly contacting colleagues, friends and family.

The frequency distribution of the contributions per player in the CrowdWater game is similar to many other citizen science projects [5,48–51]: there are a few dedicated participants who play very frequently and contribute the majority of the votes, whereas many participants have only tried the game once or play infrequently (see 3.1 Participants).

2.3 Analysis of the CrowdWater game data

2.3.1 Correction of app data through the game. The aim of the game is to check the water level class data submitted through the CrowdWater app and if necessary correct the

water level class observations. For each observation that has been classified by 15 or more players in the game, we computed the mean water level class from all the game votes. To exclude the influence of outliers, only the values within the 10^{th} and 90^{th} percentiles were used to compute the mean. This mean value differs from the current implementation of the CrowdWater game, where the approved water level class is determined based on the median value. The discrepancy between the analyses in this paper and the implementation in this game is due to the game being implemented earlier; we expect that the game will be adapted in the near future so that it also uses the mean water level class as the approved value as this more accurately reflects the correct water level near class boundaries (see results section 3.2 Vote distribution per observation and data correction). The difference between the mean water level class from the game and the water level class submitted via the app can be divided into three categories: no discrepancy, a water level class correction, and a higher water level class resolution.

- *No discrepancy*: If the mean game vote is within < 0.25 classes from the original value submitted via the app, no correction of the original app value is necessary.
- *Water level class correction*: If the mean value of the game votes is more than 0.75 water level classes away from the value submitted via the app, either the original app value or the mean game vote needs to be corrected.
- *Higher water level class resolution*: The water level of a stream can be at the border of two water level classes. In the app the citizen scientist has to decide on one of these neighbouring classes. In the game, the player also has to decide on one of the classes, but based on the distribution of the votes from many players it sometimes becomes apparent, that the actual water level class is in between the two classes. Therefore, if the mean game vote is between 0.25 and 0.75 classes away from the value submitted via the app, it could be considered as a half-class, i.e., a value with a higher resolution of the water level class scale than is possible in the app.

If the water level class submitted via the app and the mean game vote do not agree (i.e. they differ by more than 0.75 water level classes), expert judgement is needed to determine which of these values is most accurate. Experts can decide based on the pictures of the stream level that are also shown in the game whether the app value or the mean game vote is more likely to be correct. Two of the authors (Strobl and Etter) checked and classified the observations individually and discussed the pictures in case their expert judgement differed. The expert judgement was only done to evaluate the performance of the CrowdWater game, to better understand the accuracy of the output of the game for future applications. This will not be done continuously for the CrowdWater game as the number of pictures that need to be assessed would quickly become unmanageable. The categories used for expert judgment were as followed:

- The original app value was correct.
- The mean game vote was correct.
- Neither was correct, but the original app value was closer to the correct value.
- Neither was correct, but the mean game vote was closer to the correct value.
- The correct value was precisely in the middle between the original app value and the mean game vote.
- The observation should have been reported, rather than voted on by players, e.g. because there was no possibility to determine the exact value based on the picture.

2.3.2 Vote distribution per observation. For each observation that had at least 15 votes, we determined the distribution of the differences between each vote and the mean game vote (i.e. the error distribution). The distribution of the errors is an important indicator of the certainty of a mean value and showed how sure *"the crowd"* was of their collective vote. We wanted to know if the error distribution of the game votes was similar to the error distribution of the water level classes for people who see the actual stream (rather than only a picture). We therefore compared the error distribution for all observations in the game with the error distribution from a previous field study, in which 517 passers-by at ten different locations were asked to estimate the water level class of the stream by comparing the current situation with a printed copy of the reference picture with the staff gauge [45]. The error distributions for the game players and the passers-by were not normally distributed, therefore we used the Mann-Whitney test to compare the medians of the two datasets and the Pearson Chi-Squared test to compare the frequency distributions.

2.3.3 Impact of the number of votes on the calculation of the mean water level class. We wanted to determine the number of votes per observation that are needed to obtain a correct mean game vote to be able to design the CrowdWater game in such a way that it accurately classifies the observations in the most efficient way possible (i.e., to remove observations from the game when they have been classified by a sufficient number of game players, so that the effort can be directed towards classifying new observations). Currently 15 votes are needed to classify a picture but this number was merely an initial guess and could be changed based on the results of this analysis. We used bootstrapping to evaluate how the number of votes per observation affects the uncertainty of the mean game vote and thus the resulting water level classification. More precisely, we compared how the mean value of the votes for a randomly chosen subset of votes (ranging from 1 to 30 votes) for each classified picture differed from the overall mean that takes all votes into account. We then determined the number of classified pictures for which this difference was less than 0.05 and less than 0.2. This was repeated 10 000 times. The analysis was only done for classified pictures with at least 30 votes (246 observations or 11% of all observations). For this analysis we used the actual mean vote, rather than the mean within the 10^{th} to 90^{th} percentile of the votes, as the exclusion of outliers was not practical for the smaller subsets of votes.

2.3.4 Accuracy per player. We also tested whether there are differences in the abilities of the players to classify an observation correctly and if this is connected to how regularly they play the game, i.e., whether regular players are better at assigning the right water level class to an observation than novice players. Therefore, we calculated the mean accuracy per player, which is the mean of the absolute difference between the vote of the player and the mean game vote (within the 10^{th} to 90^{th} percentile of all votes) for all of their votes and subtracted this value from 10 (the maximum possible divergence). Thus an accuracy score of ten indicates a perfect score (i.e., the votes of the player were always the same as the average vote from all players), whereas a low value indicates that the votes of the player were often different from the average vote. To check whether or not the mean accuracy per player was significantly different for the regular and novice players, we used the Mann-Whitney test ($p < 0.05$) because the Shapiro-Wilk normality test indicated that the mean accuracy values were not normally distributed. Regular players were defined as players who played more than two full rounds of the game (38% of all players, representing 96% of all votes), whereas novice players played fewer rounds.

2.4 Survey

To address questions related to the motivation of the participants, we sent out a short survey to everybody who had played the CrowdWater game at least once before 11.02.2019 (145 players), using the email addresses that were used to register for the game. The questions in the

survey took 5–10 minutes to complete and covered several topics, such as what motivated the respondents to play the game, which game aspects they liked, and which ones were frustrating. In the survey we also asked if the respondents had used the CrowdWater app and how their experience with the app compared to their experience with the game. The full survey can be found in the supplementary material 1 (S1 File).

3. Results

3.1 Participants

By 28.02.2019, 153 players had registered for the CrowdWater game and contributed at least one vote; in total, 33 176 picture pairs have been compared. However, only 58 players had played more than two rounds, indicating that many participants only tried the game once or twice. The average number of observations classified per participant was 148, but the median was only 12. The mean number of classifications for the five most dedicated contributors was 1829. These results indicate a very skewed distribution of the number of classifications among the participants. Few of the participants who participated in the survey had watched the tutorial movie (36%) but more participants read the manual (61%) before playing the game for the first time.

3.2 Vote distribution per observation and data correction

The agreement of votes for classified pictures varied significantly from observation to observation and sometimes even for observations taken at different times for the same site. However, the app values and the mean game vote rarely differed by more than one water level class. For 43% of all classified observations there was no difference between the original water level class submitted via the app and the mean game vote, meaning that the app user and the mean vote of the game players agreed. For 27% of all observations, the mean value from the game differed by half a water level class, which should not be considered an error, but rather an increase in the resolution of the data (i.e. indicating a water level between two class boundaries). For 30% of all classified observations the mean game vote and the app entry differed by at least one class. For 20% of all classified observations the disagreement was exactly one class; for only 10% of the classified observations, the mean game vote and the original app value differed by more than one class (Fig 3).

The agreement among the game players was particularly high for observations for which the water level was relatively similar to that in the reference picture (i.e., the mean vote had a water level class of zero; Fig 4).

Based on the mean game value, 390 of all 846 classified observations (46%) fell into water level class 0 (i.e., similar to the water level class as in the reference picture), whereas all other classes had $< 10\%$ of the observations respectively. For 263 (31%) classified observations the mean vote indicated a half-class (Fig 5).

For 77% of the observations with a water level class of +1, the mean vote was class +1 ($n = 64$), whereas 68% of the observations with a water level class of -1 had a mean vote class -1 ($n = 45$).

The frequency distribution of the differences in the votes per observation from the mean game vote for that observation was similar to the distribution of the errors in the water level class assignment for 517 passers-by who estimated the water level for ten streams with the same virtual staff gauge approach [45]. For both the game and the real life situation more than 48% of the participants chose the right class, and less than 3% were more than two classes off. The median difference in the water level class values (0 for the game and the passers-by) and the frequency distribution were not significantly different either ($p < 0.05$; Fig 6). The accuracy was comparable, as the two distributions were not significantly different from each other (based on a Pearson Chi-Squared test, $p < 0.05$).

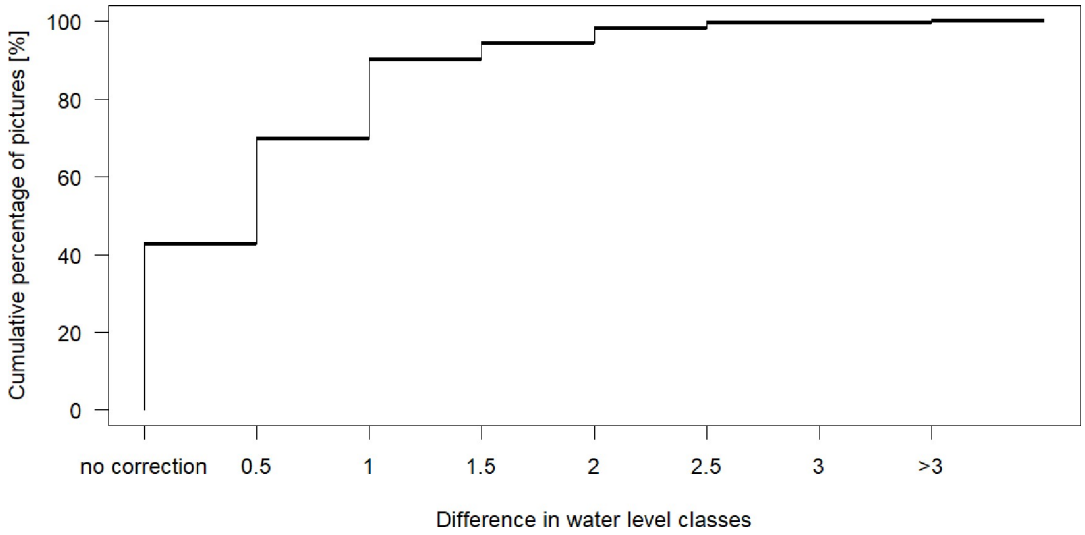


Fig 3. Cumulative frequency distribution of all corrections. Corrections of the original app value based on the mean game votes (between the 10th and 90th percentile). 100% = 841 classified observations.

https://doi.org/10.1371/journal.pone.0222579.g003

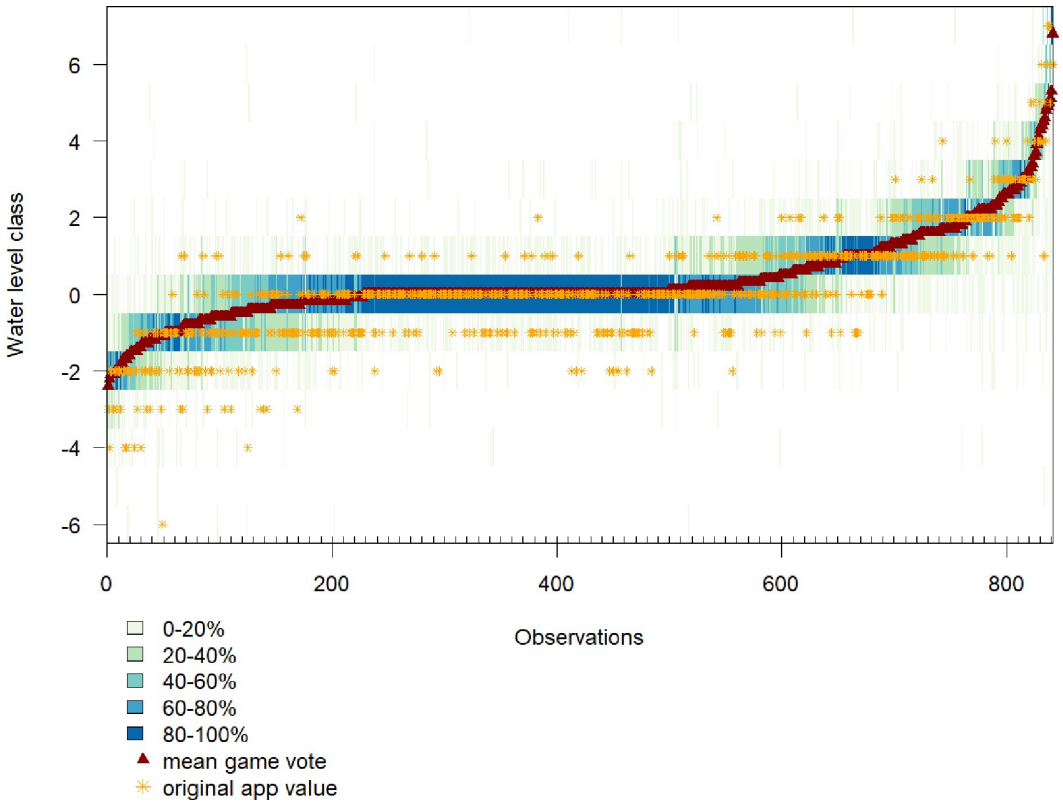


Fig 4. Agreement among players (in %) per classified observation. Each column represents one observation, sorted according to the mean game vote (red triangle). Darker colours represent a higher agreement and lighter colours a lower agreement among the players. The original value of the water level class submitted via the app is indicated by the orange star.

https://doi.org/10.1371/journal.pone.0222579.g004

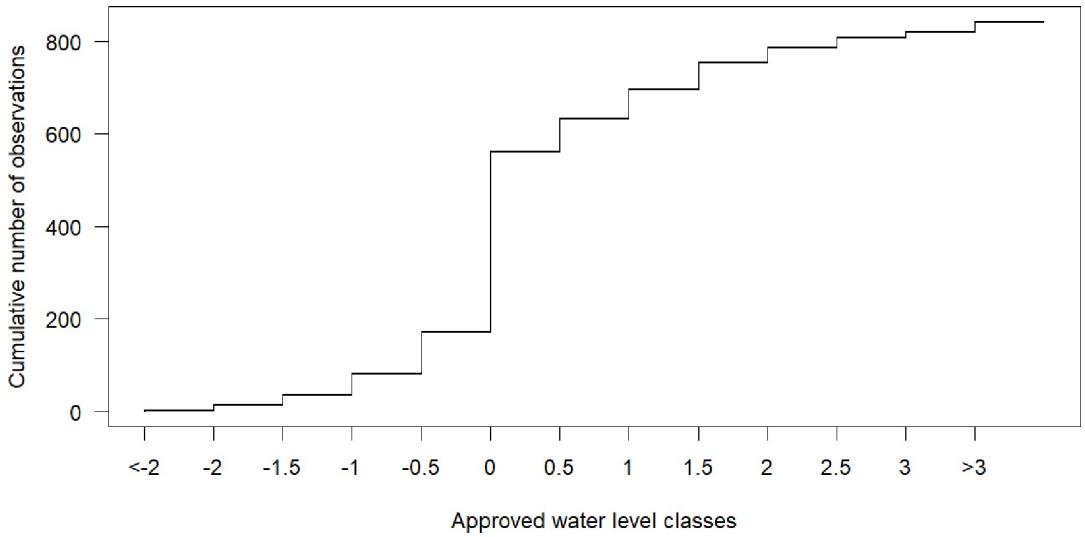


Fig 5. Cumulative number of classified observations per water level class based on the mean game value.

https://doi.org/10.1371/journal.pone.0222579.g005

All 252 observations for which the mean game vote and the original app value differed by one or more class, were classified through expert judgement. For 60% of these cases the mean game vote was considered to be correct and for 14% of cases neither the mean game vote nor the app value was correct, but the mean game vote was closer to the correct value. For 8% of these cases the app value was correct and in 1% of cases neither were correct, but the app value was closer to the correct value. For 8% of cases the correct value was precisely between the app value and the mean game vote. For 9% of the cases the observations should have been reported, rather than getting a water level class vote.

3.3 Accuracy per player

The median of the mean accuracy per player for the 58 regular players (i.e. $>$ 24 classifications per player) (9.60) was significantly better than the median of the mean accuracy per player for the 94 novice players ($\leq$ 24 classifications per player) (9.26). The range in the mean accuracy per player was smaller for the regular players as well (1.24 classes for the regular players vs. 3.63 classes for the novice players). Players seem to get even better with more rounds, as the median for players with more than four rounds (9.62) was also significantly better than for players between two and four rounds (9.53). All very frequent players (those who classified $>$ 500 observations; n = 11) had a mean accuracy that was higher (i.e., better) than the median of the mean accuracy for all players (Fig 7), however statistical differences could not be calculated due to a small sample size of very frequent players.

3.4 Observation reports

Almost half of the players used the report function at least once. Of all players who reported at least one observation, 77% were regular players, or, expressed differently 35% of the regular players and 87% of the novice players never reported an observation.

The report function was used at least once for 8% of all observations (classified and unclassified) that were in the game (n = 193). After an observation has received 15 reports, it is removed from the game. So far this occurred for only three observations (Fig 8).

The most common reason for reporting an observation was *"other reason"* (47% of all reports). Within this category, 48% of all reports stated that the picture was too dark. The

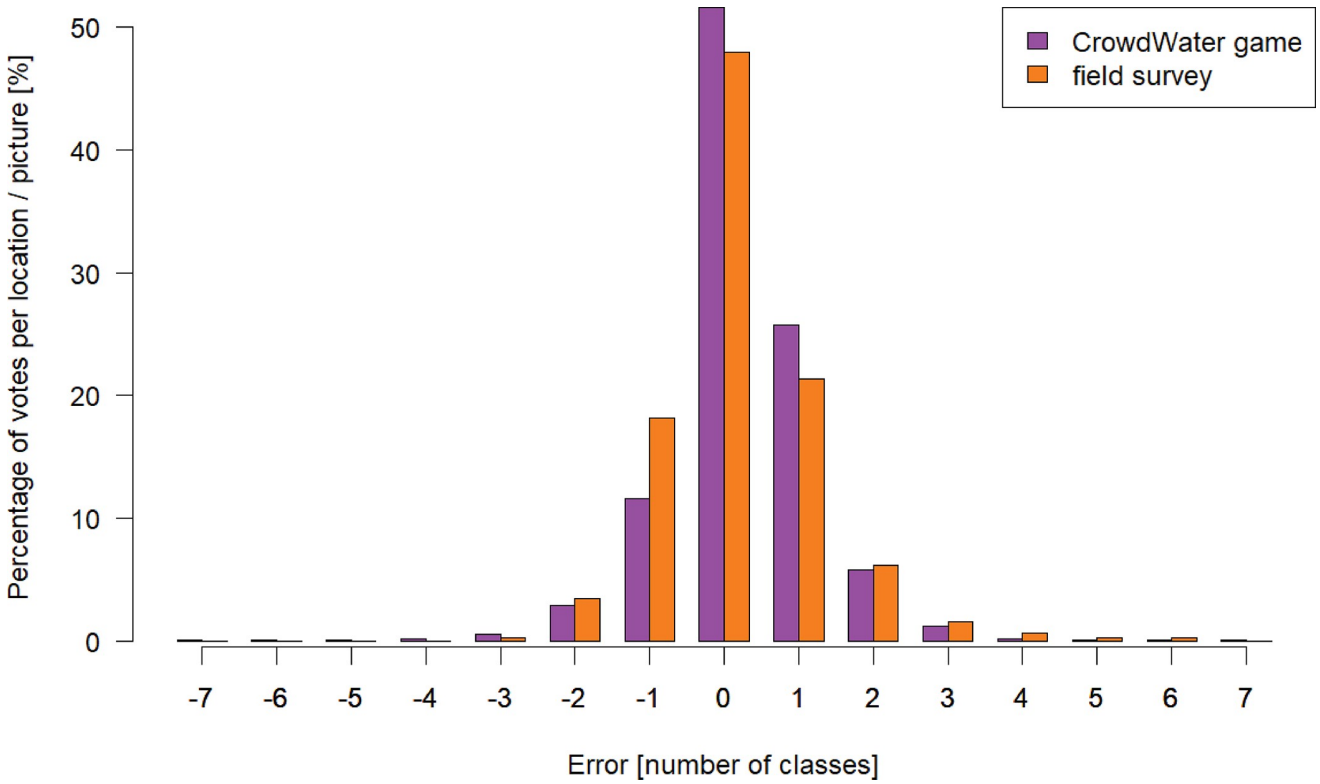


Fig 6. Error distribution of the water level classes for the CrowdWater game and a field survey. Comparison of the error distribution of the votes in the CrowdWater game (difference between the vote of a player and the mean game vote for that observation; n = 841) and for previously held field surveys (n = 517) [data from 49].

https://doi.org/10.1371/journal.pone.0222579.g006

second most common report was *"the location has changed and the reference image is unrecognizable"* (32%), followed by *"The staff gauge is not placed correctly"* (10%) and *"The staff gauge is missing"* (10%). The other report categories were rarely used (< 3% each) (Table 1).

3.5 Impact of the number of votes on the mean game vote

The impact of the number of votes on the mean of the game votes for that observation shows at what point the mean game vote becomes stable, i.e., the mean value does not change with additional votes. The results of the bootstrapping analyses indicate that for 89% of observations the error was ≤ 0.2 after 15 votes, for 90% of observations the error was ≤ 0.2 after 16 votes and for 95% of the observations after 20 votes (median values for all 10 000 iterations) (Fig 9). An error ≤ 0.2 would still be rounded to the approved water level class. More votes steadily increased the percentage of observations above these thresholds.

3.6 Survey

A quarter of all players [36] who were sent the link to the survey filled in the survey. Half of the respondents of the survey played the CrowdWater game at least once a week, a quarter of the respondents played the game one to three times per month, and another quarter of the respondents had played only once.

The main motivations for playing the game were enjoyment in playing the game, a general interest in hydrology, and being part of the CrowdWater community. Contributing to science,

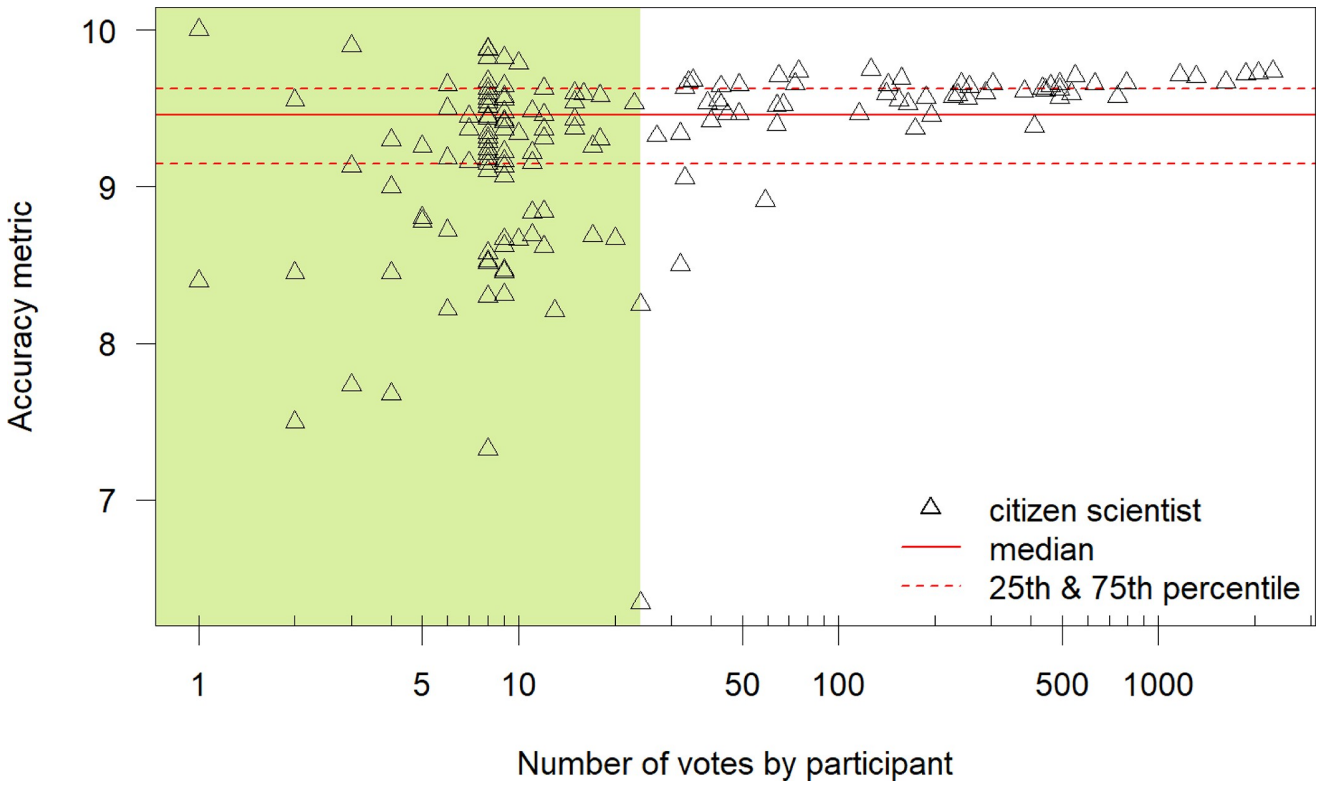


Fig 7. Mean accuracy per player. Mean accuracy per player as a function of the number of observations that that player classified (each triangle represents one player). The lines indicate the median accuracy for all players (solid line) and the 25th and 75th percentile (dashed lines). The green shading indicates the novice players who played a maximum of two rounds (24 classifications). Note the log scale on the x-axis.

https://doi.org/10.1371/journal.pone.0222579.g007

helping the environment, and the monthly competitions were less frequently mentioned as motivating factors (Fig 10). The majority of respondents stated, that they enjoy *"classifying difficult pictures, even though I might not get full points"* compared to a minority who said that they enjoyed *"classifying easy pictures, because I will likely get full points"*. Almost two thirds of respondents said that they enjoyed *"competing against each other"*; one person found that *"the points and competition are unnecessary and distract from the scientific goal"*. Overall the three aspects found to be most frustrating and marked by about half of all respondents were *"difficulty finding adequate references in the pictures"*, *"pictures that are not taken from the same angle"* and *"not getting full points, even though I am sure of my vote"*.

The majority (67%) of the players who responded to the survey also use the CrowdWater app. Almost half of them enjoyed both activities equally, a third enjoyed using the app more, while a quarter enjoyed the game more. The large majority (79%) of respondents who also use the app indicated that the game helped them *"be more aware of how to place a staff gauge in the app"*. Half of the respondents stated that the game helped them *"to estimate water level classes in the app"*.

4. Discussion

4.1 Can peer-review improve the quality of crowdsourced water level class data?

Many publications have reported on the accuracy of citizen science games, with game aims, topics and styles covering an extensive range. Therefore the results and the accuracy metrics

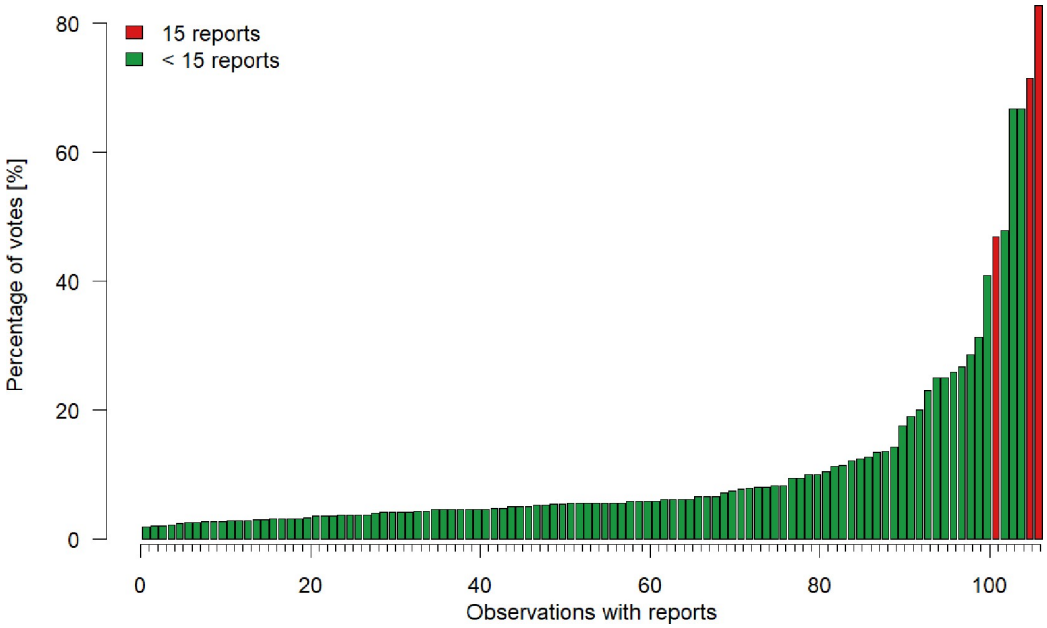


Fig 8. Percentage of reports over all votes. Reports for each of the 841 classified observations (i.e. at least 15 votes in total) that received at least one report. The red colour indicates that the observation received 15 reports, after which it is excluded from the game.

https://doi.org/10.1371/journal.pone.0222579.g008

used in these previous publications vary considerably. Generally the conclusions have been very positive and many of these citizen science games are still online and collect valuable scientific data [24,35,37,49,52,53]. The results of the CrowdWater game are also positive and suggest that showing the same picture pair to multiple players and taking their collective vote as the approved value can help improve the quality of the water level class data that are collected by the CrowdWater app. A big benefit of peer-review compared to a data filter, is that incorrect data are not only filtered out but can be updated and therefore can still be used as a valid data point in later analyses. Furthermore, it does not require pre-defined criteria of what data are likely to be incorrect.

For 70% of all water level class observations, there was either no difference or a difference of only half a class between the mean game vote and the original app value. For 74% of the observations for which there was a difference between the mean game vote and the original app value, the game provided a more accurate estimate of the water level class than the app. This suggests that the game is a valuable tool to check the quality of the water level data provided through the app. We would therefore strongly recommend collecting pictures together with any citizen science data (if feasible), even in projects where this does not seem essential at first.

Table 1. Reasons given for a report as a percentage of the overall number of reports.

Report reason	Percentage of reports
The photo is ok, but I don't know the category.	1%
The staff gauge is not placed correctly.	9%
The staff gauge is missing.	9%
The approved value is clearly incorrect.	2%
The location has changed and the reference image is unrecognizable.	32%
Other reason: too dark to classify	23%
Other reason: all other reasons	24%

https://doi.org/10.1371/journal.pone.0222579.t001

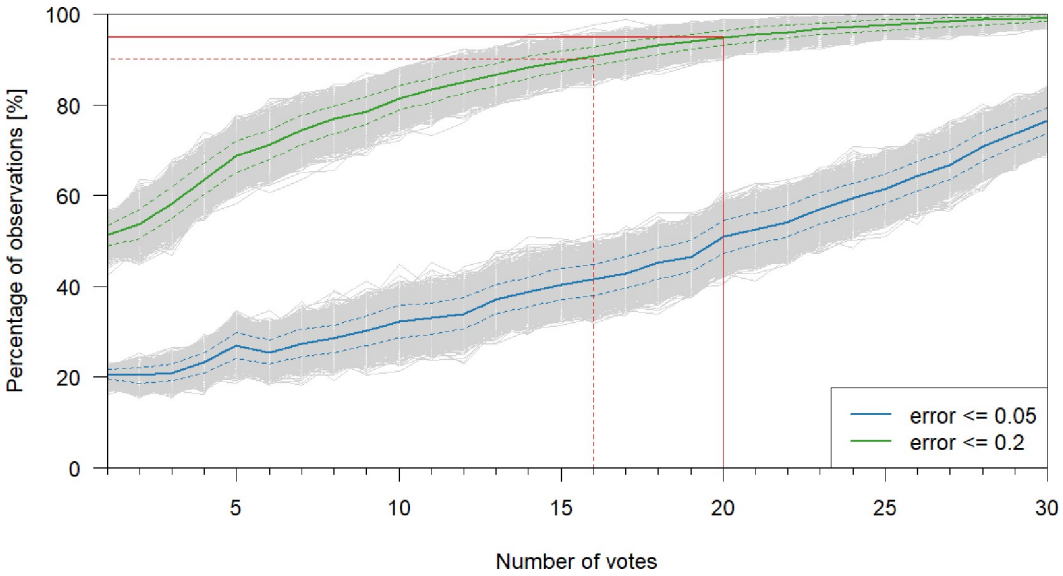


Fig 9. Impact of number of votes. Percentage of observations with an error ≤ 0.05 (median in blue) or an error ≤ 0.2 (median in green) as a function of the number of votes per observation. The dashed blue and green lines indicate the 10^{th} and 90^{th} percentile, while the grey lines show the results for all 10 000 iterations. The solid red lines indicate that for 90% of observations the error is ≤ 0.2 after 16 votes. The dashed red lines indicate that for 95% of observations the error is ≤ 0.2 after 20 votes.

https://doi.org/10.1371/journal.pone.0222579.g009

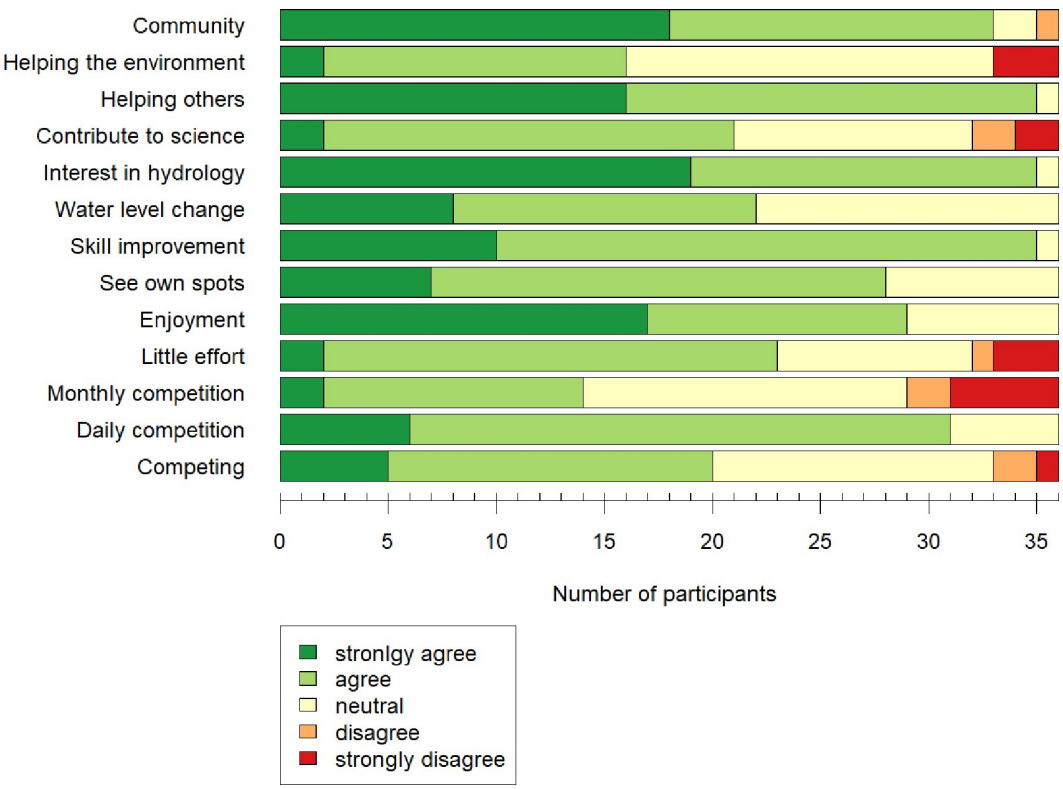


Fig 10. Motivation of the participants to play the CrowdWater game. The survey questions can be found in supplementary material 1 (S1 File).

https://doi.org/10.1371/journal.pone.0222579.g010

The frequency distribution of the vote errors in the field survey was similar to that of the CrowdWater game, even though the reference pictures in the field survey were all initiated by us (which should have guaranteed a reasonable quality) and the participants were able to look at the stream directly. In the game, the reference pictures are created by citizen scientists using the app (where sub-optimal reference pictures do occur) and the player can only take a look at a picture of the stream. The sense of depth may be different for the picture than when one sees the actual stream. In spite of these differences, and even though not all pictures are optimally sharp or free from distortions, citizen scientists were just as good at determining the water level class in the game as the passers-by that assessed the water level class outside. This furthermore suggests that it is beneficial to obtain pictures with the submission of the data via the app and to use these pictures in a gamified approach to improve data quality.

However, in some cases (9% of all observations for which the mean game vote differed from the water level class submitted via the app), the water level class obtained from the mean game vote was less accurate than the original app value (as determined by expert checks). When the number of observations increases, it will be impossible to determine whether the mean game vote or the original app value is right. It would be useful to automatically identify these cases, to avoid changing a correct app value. To identify these observations, indicators could be used, such as kurtosis, as a proxy for the vote distribution, bimodality to indicate diverging interpretations of the same picture, location to indicate a poorly placed staff gauge in the reference picture or the overall contribution of the app user or mean accuracy of the contributing game players [54]. Currently not enough data are available to assess, whether these indicators can be reliably used, however, in the future this topic might be worth revisiting. Another option to assess the correctness of the original water level class submitted via the app would be to go back to the observer who submitted the original via the app and to ask how certain (s)he is about the initial estimate. In some cases, such as accidental misclassification, the discrepancy might be easily resolved. Additionally this would provide feedback to app users, which might provide training and increase the accuracy of their data. By receiving feedback app users find out that their data are checked. This could increase their confidence in being able to contribute high-quality data, which was discovered to be useful in a study where citizen scientists from seven different online citizen science projects were interviewed [5]. If the app user insists that his/her water level class estimate is correct, the observation could be reviewed by an expert, as this would likely occur for fewer cases than all corrections. Currently, without any further means of validating the game results, it still makes sense to use the mean game vote as the approved value, as in 74% of cases the game provided a more accurate water level class than the original submission via the app.

4.2 Can the vote distribution per observation provide additional information?

The primary reason for using several votes per observation was to minimise errors in the water level class assignment because the errors of individual game players likely average out (i.e. wisdom of the crowd [55]). The vote distribution per observation shows if there is a high agreement between players, which suggests a higher certainty for the resulting mean game vote, and thus higher certainty for the water level class. The results showed that the highest agreement between players is at class zero, which per definition represents a situation that closely resembles the reference picture. Some of the uncertainty for very high or low water level observations can therefore partly be derived from the interpretations that game players make to assess a very different looking picture. The interpretation is particularly tricky, if the correct water level class is below the water surface in the reference picture, as the relevant reference features might not be visible.

This is supported by the results indicating a higher percentage of votes for the approved class for observations with class plus one compared to class minus one (77% vs. 68%).

The vote distribution per observation can also indicate if the water level is at the boundary between two classes. This results in higher resolution water level class data than is possible in the app. The app user has to decide on one of the two classes, even if the user believes that the water level is exactly at the border of two classes. While the individual game users also had to decide on either the upper or the lower class, their vote distribution could show that the water level was likely in the middle of two classes. This higher resolution water level class data may be useful for further analyses of these data, e.g. to calibrate a hydrological model.

For a few observations, we detected a bimodal distribution in the game votes (not between two neighbouring classes), which could indicate that a picture was unclear and allowed for multiple interpretations. These observations all came from the same location, which had large changes in the amount of sediment in the stream bed after the reference picture was taken (both deposition and scour). Based on this small sample, it is not possible to conclude that a bimodal distribution only occurs for such scenarios and can be used to filter out such situations. However, similar results regarding the agreement among players were found in a study based on cropland identification, where *"crowdsourcing appeared particularly cost-effective in areas that were easy to interpret and allowed difficult or problematic sampling units to be identified, i.e., as evidenced by a lack of consensus between volunteers."* [54].

4.3 Are votes from regular players more accurate?

The votes from regular players, were statistically significantly more accurate than the votes from novice players. There are two possible reasons for this higher accuracy. On the one hand players might get better after playing several rounds and therefore their mean accuracy improves. On the other hand, novice players might notice that they consistently get few or no points and therefore drop out of the game. The survey showed that only a minority of the players really cared about achieving a high score, but nonetheless some stated that they found it frustrating to vote for difficult pictures, as they might not get full points. Perhaps the players who got few points and dropped out of the game after a few rounds were the same ones that were eager to get many points and win, however, we do not have data to check this assumption.

A similar difference in accuracy for regular and novice players was not observed for the Forgotten Island and Happy Match games, and only a minimal difference was found for the game Happy Moths [49]. However, for the Cropland Capture game, it was shown that the score of players could indeed improve over time [56]. This suggests that the difference in the accuracy of the votes between regular players and novice players and the potential for improvement of the accuracy when playing the game depends on the game.

There is also the possibility of bias in the game. Observations had to be classified by 15 different players before they were considered classified. Due to game logistics, further classifications were not necessarily done by different players. Therefore, it is possible for the same player to classify an observation more than once, thereby influencing the mean game vote which results in a higher accuracy. However, due to a large number of players and many observations available in the game, we assume that this effect is negligible.

4.4 Can players correctly identify unsuitable observations through the report function?

The report function is a valuable tool to filter out unsuitable observations. When an observation is reported because *"The staff gauge is not placed correctly."*, *"The staff gauge is missing."* and *"The location has changed and the reference image is unrecognizable."*, the observation

should not only be removed from the game, but also from the app, as e.g. a reference picture without a staff gauge cannot be used to obtain a water level class time series. For the categories *"The photo is ok, but I don't know the category."*, *"The correct value is clearly incorrect."* and *"Other reason"* the observation and picture might not be suitable for the game, but might still be a valuable data point. The reported reason *"the picture is too dark"* is an example of a picture being unsuitable for the game, but the app player might still have seen the location adequately to estimate the water level class correctly. On the other hand, the report reason *"The location has changed and the reference image is unrecognizable."* indicates a problem with the reference picture or it might be challenging to find a suitable reference. Here the number of votes could be used as an indication: if the majority of players reported the observation it is likely that there was an issue with the reference picture, but if only a few players reported the observation it might simply be difficult to find a suitable reference. The observation in the app should be removed, if the streambed has indeed changed significantly between the time that the reference picture was taken and the new observation.

The fact that roughly half of all players never used the report function, could indicate that these players are unaware of the report button or are unsure when to use it. This is also supported by the fact that only very few novice players (13%) used the report function at least once (compared to 65% of regular players). The infrequent use of the report function was also reported for the Cropland Capture game, which has a button so that players can choose *"maybe"* instead of stating whether a picture displays cropland or not. This button was only used infrequently (rarely over 50% of the votes per picture), even for pictures that were difficult to classify [34].

If a large number of players cannot find the report button in the CrowdWater game, it may take longer for certain spots to be removed from the game. There were several ambiguous cases, when it might still be possible to guess the water level class on a relatively dark picture, but the estimate was likely uncertain. In these cases some users opted for the report button and others decided to vote for the most likely water level class. In addition, there was a chance that players just guessed that the water level class is zero (which is the most frequently occurring water level class) in order to get full points.

4.5 How many votes are necessary to achieve a stable mean?

Our current number of 15 votes per observation to consider an observation classified and keeping it in the game for 100 votes seems to work well because for 90% of the observations the error was ≤ 0.2 after 16 votes. Allocating game points with a certainty of just under 90% seems sufficiently accurate, especially as more votes are collected overall. We will therefore leave the classification threshold at 15 votes for the point allocation, but will reduce the total number of votes per observation to 50, in order to more quickly complete the classification of observations within the game to more quickly classify observations.

Other projects have investigated the ideal number of votes in a similar way. The cut-off number for votes varies depending on the citizen science project but is comparable to the cut-off value found for the CrowdWater game. The Cyclone Center decided on ten classifications per picture to reach a *"statistically reasonable consensus"* [22], the project Pattern Perception used a *"retirement limit"* of 20 votes, in OpenStreetMap 15 contributors per square kilometre resulted in a very good positional accuracy [20] and in the MalariaSpot game 22 votes from non-expert players or 13 votes from trained players resulted in an accuracy higher than 99% [57]. StallCatchers tried to reduce this number through individually weighed sensitivity measures in order to quickly advance the study field and to ensure that the time of the citizens is spent efficiently [58,59]. StallCatchers ultimately arrived at a flexible number of necessary

votes per picture based on voting consistency and user experience, with an average of seven votes [58,59]. Such methods could in the future also be implemented in the CrowdWater game, to keep the number of necessary votes per observation flexible, e.g. by checking the voting agreement for each observation and by taking the accuracy of the player into account. For example, an observation, where five usually rather accurate players are in full agreement, is likely to already yield the correct mean value, whereas high disagreement for an observation might require more votes to obtain an accurate classification.

4.6 What motivates participants to play the CrowdWater game?

The survey results indicate that the majority of the game players enjoy *"helping others"*, the game in general, or are interested in hydrology. The majority (75%) of the players enjoy classifying difficult pictures, even though they might get fewer points, which also suggests that most users like the game for its purpose, rather than for the gamified aspects. This is in contradiction with the statement that they found it frustrating not to get full points, even when they were sure about their vote. Only a minority of players mentioned that they enjoy the competition but one player indicated that s(he) found the gamified aspects distracting. Based on the survey results, we decided to keep the gamified aspects as they are currently implemented.

The survey also showed that there was a large overlap between the game players and app contributors. This was not what we had expected, as we thought that the different approaches would appeal to different people. However, many of the people we could initially reach with news about the CrowdWater game were already interested in the CrowdWater project. Perhaps in the future, when both parts of the project are better known, the two user groups may become more distinct. One advantage of the overlap between the game players and the app users is that 79% of the game players who participated in the survey indicated that playing the game made them more aware of how to place a staff gauge in the app. This is likely because the players see examples of both good and poorly placed staff gauges in the game, which might make them more aware of how they can do it better themselves. This means that the game could be used as initial training before using the app, as some of the reference pictures do not have the correct angle, size or sometimes lack a staff gauge altogether [39].

4.7 Further research

The CrowdWater game can improve the accuracy of the water level class data gathered through the CrowdWater app and thereby enhanced the usability of the data, e.g. for hydrological modelling. Future research will have to investigate how to best incorporate the game results into the app. One method could be to automatically update the app data with the values derived from the game. However, this could also put errors into the app data that were not there before. Alternatively, deviations between the app and game values could be flagged, so that an expert can look at these particular data points. The main issue with the second approach is the scalability, as the database may quickly become too large for such an approach. However, super users who have played more than a certain number of rounds of the game may be involved in this as well, as the accuracy of their data was very high and their votes could potentially be weighed more. The original app value could also be taken into account by simply adding it to the game votes as an additional vote, perhaps with additional weight as the app user gets to see the actual location, whereas the game players only see the picture.

Another interesting topic to be investigated in the future, is that of automating the CrowdWater game. Michelucci and Dickinson [60] defined the phrase *"human computation"* as the *"combination of humans and computers to accomplish a task that neither can do alone"*. This is also reflected in Kawrykow et al. [35] who say that *"crowdsourcing begins where automation*

fails". The boundaries of what a computer can accomplish are however likely to shift in the future, which means that it is feasible that some of the steps that are currently done by game players, such as an automatic recognition of the virtual staff gauge and water level by a computer, could be outsourced to computers. At that stage it might still be possible to continue the game, but to adjust the specific tasks to what is needed. This balance will always have to be reassessed carefully, as *"intuition and reasoning often make humans more effective than computer algorithms in various realms of problem solving."* [21]. Additionally people are better at visual classifications, as shown in projects such as StallCatchers [58,59], Galaxy Zoo [23], Snapshot Serengeti [24] and Cyclone center [22].

5. Conclusions

The CrowdWater game allows checking and correcting crowdsourced water level class data based on the pictures that are submitted by citizen scientists through the CrowdWater app. This means that both data submission and data quality control are crowdsourced, which provides two different tasks (one in the field and one online) for citizen scientists who want to join the CrowdWater project. The results of this study indicate that the CrowdWater game improves the accuracy of the water level class data that are collected via the app by correcting a third of all app values. This improves water level class data for future purposes such as hydrological modelling. The game also helped to increase the resolution of the data as a third of all classified pictures had a mean game vote that fell into a half class. This provides higher resolution data than is currently possible through the app. The game can also be used to filter unusable observations or reference pictures through the report function, e.g. if the virtual staff gauge was not placed correctly, but half of the players, including a quarter of the regular players, never used this function.

Through the pictures provided via the app, the game can ensure data quality control for time-series of water level class data obtained via citizen science. We, therefore, recommend that citizen science projects obtain pictures, in addition to an observation value, so that they can be used for data quality control. While other citizen science games so far have mostly used professional pictures, this study shows that games can also be based on crowdsourced pictures of environmental observations. Additionally, we recommend that citizen science games aim for regular contributors through suitable advertisement and achievable daily goals, as regular players tend to have a higher voting accuracy. Games or citizen science projects should also determine early during a project the right numbers of votes per observation, as this has the potential to save time and effort of project organisers and citizen scientists.

Even though the results show that the majority of the submitted crowdsourced water level class data is correct, even without quality control, the results of this study indicate that the CrowdWater game improves the data by correcting water level class observations, increasing the data resolution and removing unusable reference pictures and observations. The results of the CrowdWater game show the potential of gamified approaches to crowdsource data quality control in citizen science projects. This is particularly valuable for variables that can change rapidly, such as water levels in our case, because other forms of data quality control are difficult because observations by different citizen scientists at the same time and place are not realistic in practice.

Supporting information

S1 File. Questions for the online survey for CrowdWater game players.
(DOCX)

Acknowledgments

We thank all 153 CrowdWater game players (as of 28.02.2019) for their time and interest in this research project and for sharing their water level class estimates with us. We also thank all CrowdWater app contributors for providing the database for the CrowdWater game and for submitting valuable hydrological data. All CrowdWater participants agreed to the terms of use (https://www.spotteron.net/terms-of-use) when first using the app or playing the game. We hope that many of the CrowdWater participants, both in the app and in the game, will continue to participate in the CrowdWater project and invite all readers to use the app and/or play the game as well.

Author Contributions

Conceptualization: Barbara Strobl, Simon Etter, Ilja van Meerveld, Jan Seibert.

Data curation: Barbara Strobl.

Formal analysis: Barbara Strobl, Simon Etter.

Funding acquisition: Ilja van Meerveld, Jan Seibert.

Methodology: Barbara Strobl, Simon Etter, Ilja van Meerveld, Jan Seibert.

Project administration: Barbara Strobl.

Supervision: Ilja van Meerveld, Jan Seibert.

Visualization: Barbara Strobl.

Writing – original draft: Barbara Strobl.

Writing – review & editing: Simon Etter, Ilja van Meerveld, Jan Seibert.

References

1. Engel SR, Voshell JR. Volunteer biological monitoring: can it accurately assess the ecological condition of streams? Am Entomol. 2002; 48:164–77.
2. Cooper CB, Shirk J, Zuckerberg B. The invisible prevalence of citizen science in global research: Migratory birds and climate change. PLoS One. 2014; 9(9).
3. Danielsen F, Jensen PM, Burgess ND, Altamirano R, Alviola PA, Andrianandrasana H, et al. A Multicountry Assessment of Tropical Resource Monitoring by Local Communities. Bioscience. 2014; 64 (3):236–51.
4. Freitag A, Meyer R, Whiteman L. Correction: Strategies Employed by Citizen Science Programs to Increase the Credibility of Their Data. Citiz Sci Theory Pract. 2016; 1(1):2.
5. Jennett C, Kloetzer L, Schneider D, Iacovides I, Cox AL, Gold M, et al. Motivations, learning and creativity in online citizen science. J Sci Commun. 2016; 15(3):1–23.
6. Bonter DN, Cooper CB. Data validation in citizen science: A case study from Project FeederWatch. Front Ecol Environ. 2012; 10(6):305–7.
7. Goodchild MF, Li L. Assuring the quality of volunteered geographic information. Spat Stat. 2012; 1:110–20.
8. Wiggins A, Newman G, Stevenson RD, Crowston K. Mechanisms for data quality and validation in citizen science. In: Seventh IEEE International Conference on e-Science Workshops. Stockholm: IEEE; 2011. p. 14–9.
9. Cohn JP. Citizen Science: Can Volunteers Do Real Research? Bioscience. 2008 Mar; 58(3):192–7.
10. Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, Rosenberg K V., et al. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. Bioscience. 2009 Dec; 59 (11):977–84.
11. Hochachka WM, Fink D, Hutchinson R a, Sheldon D, Wong W-K, Kelling S. Data-intensive science applied to broad-scale citizen science. Trends Ecol Evol. 2012 Feb; 27(2):130–7. https://doi.org/10.1016/j.tree.2011.11.006 PMID: 22192976

12. Walker D, Forsythe N, Parkin G, Gowing J. Filling the observational void: Scientific value and quantitative validation of hydrometeorological data from a community-based monitoring programme. J Hydrol. 2016 Jul; 538:713–25.
13. Wiggins A, Crowston K. From Conservation to Crowdsourcing: A Typology of Citizen Science. In: 2011 44th Hawaii International Conference on System Sciences. IEEE; 2011. p. 1–10.
14. Yu J, Kelling S, Gerbracht J, Wong W-K. Automated data verification in a large-scale citizen science project: A case study. 2012 IEEE 8th Int Conf E-Science. 2012;1–8.
15. Bird TJ, Bates AE, Lefcheck JS, Hill N a., Thomson RJ, Edgar GJ, et al. Statistical solutions for error and bias in global citizen science datasets. Biol Conserv. 2014; 173:144–54.
16. Edgar GJ, Barrett NS, Morton AJ. Biases associated with the use of underwater visual census techniques to quantify the density and size-structure of fish populations. J Exp Mar Bio Ecol. 2004; 308 (2):269–90.
17. See L, Sturn T, Perger C, Fritz S, Mccallum I, Salk C. Cropland Capture: A Gaming Approach to Improve Global Land Cover. In: Huerta, Schade, Granell, editors. Connecting a Digital Europe Through Location and Place Proceedings of the AGILE 2014 International Conference on Geographic Information Science. Castellón; 2014. p. 3–6.
18. Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, Kelling S. eBird: A citizen-based bird observation network in the biological sciences. Biol Conserv. 2009 Oct; 142(10):2282–92.
19. Foody GM, See L, Fritz S, Van der Velde M, Perger C, Schill C, et al. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. Trans GIS. 2013; 17(6):847–60.
20. Haklay M (Muki), Basiouka S, Antoniou V, Ather A. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. Cartogr J. 2010; 47(4):315–22.
21. Koch J, Stisen S. Citizen science: A new perspective to advance spatial pattern evaluation in hydrology. PLoS One. 2017; 12(5):1–20.
22. Hennon CC, Knapp KR, Schreck CJ III, Stevens SE, Kossin JP, Thorne PW, et al. Cyclone center: can citizen scientists improve tropical cyclone intensity records? Bull Am Meteorol Soc. 2015; 96(4):591–608.
23. Lintott CJ, Schawinski K, Slosar A, Land K, Bamford S, Thomas D, et al. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Mon Not R Astron Soc. 2008; 389(3):1179–89.
24. Swanson A, Kosmala M, Lintott C, Packer C. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conserv Biol. 2016; 30(3):520–31. https://doi.org/10.1111/cobi.12695 PMID: 27111678
25. MacDonald EA, Case NA, Clayton JH, Hall MK, Heavner M, Lalone N, et al. Aurorasaurus: A citizen science platform for viewing and reporting the aurora. Sp Weather. 2015; 13(9):548–59.
26. Franzoni C, Sauermann H. Crowd science: The organization of scientific research in open collaborative projects. Res Policy. 2014 Feb; 43(1):1–20.
27. Iacovides I, Jennett C, Cornish-Trestrail C, Cox AL. Do games attract or sustain engagement in citizen science? CHI '13 Ext Abstr Hum Factors Comput Syst—CHI EA '13. 2013; 1101.
28. Prestopnik NR, Tang J. Points, stories, worlds, and diegesis: Comparing player experiences in two citizen science games. Comput Human Behav. 2015; 52:492–506.
29. Baaden M, Delalande O, Ferey N, Pasquali S, Waldispühl J, Taly A. Ten simple rules to create a serious game, illustrated with examples from structural biology. PLoS Comput Biol. 2018; 14(3):1–9.
30. Ponti M, Hillman T, Kullenberg C, Kasperowski D. Getting it Right or Being Top Rank: Games in Citizen Science. Citiz Sci Theory Pract. 2018; 3(1):1–12.
31. Schrier K. Knowledge games. Baltimore: Johns Hopkins University Press; 2016. 270 p.
32. Law E, Ahn L von. Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games. In: Proc CHI'09. 2009. p. 1197–206.
33. von Ahn L. Games with a Purpose. Computer (Long Beach Calif). 2006; 39(6):92–4.
34. Salk CF, Sturn T, See L, Fritz S, Perger C. Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game. Int J Digit Earth. 2016; 9(4):410–26.
35. Kawrykow A, Roumanis G, Kam A, Kwak D, Leung C, Wu C, et al. Phylo: A citizen science approach for improving multiple sequence alignment. PLoS One. 2012; 7(3).
36. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, et al. Predicting protein structures with a multiplayer online game. Nature. 2010 Aug 5; 466(7307):756–60. https://doi.org/10.1038/nature09304 PMID: 20686574

37. Horowitz S, Koepnick B, Martin R, Tymieniecki A, Winburn AA, Cooper S, et al. Determining crystal structures through crowdsourcing and coursework. Nat Commun. 2016; 7.
38. Curtis V. Online citizen science games: Opportunities for the biological sciences. Appl Transl Genomics. 2014; 3(4):90–4.
39. Seibert J, Strobl B, Etter S, Hummer P, van Meerveld HJ. Virtual Staff Gauges for Crowd-Based Stream Level Observations. Front Earth Sci. 2019 Apr 12;7.
40. Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, et al. The biodiversity of species and their rates of extinction, distribution, and protection. Science (80-). 2014; 344(6187).
41. Silvertown J, Harvey M, Greenwood R, Dodd M, Rosewell J, Rebelo T, et al. Crowdsourcing the identification of organisms: A case-study of iSpot. Zookeys. 2015; 480:125–46.
42. Kampf S, Strobl B, Hammond J, Annenberg A, Etter S, Martin C, et al. Testing the waters: Mobile apps for crowdsourced streamflow data. Eos (Washington DC). 2018; 99.
43. Seibert J, van Meerveld HJ, Etter S, Strobl B, Assendelft R, Hummer P. Wasserdaten sammeln mit dem Smartphone–Wie können Menschen messen, was hydrologische Modelle brauchen? Hydrol und Wasserbewirtschaftung. 2019; 63(2).
44. Etter S, Strobl B, Seibert J, van Meerveld I. Value of uncertain streamflow observations for hydrological modelling. Hydrol Earth Syst Sci. 2018; 22:5243–57.
45. Strobl B, Etter S, van Meerveld I, Seibert J. Accuracy of crowdsourced streamflow and stream level class estimates. Hydrol Sci J. 2019 Mar 29;(Special issue on hydrological data: opportunities and barriers):1–19.
46. Crowston K, Prestopnik NR. Motivation and data quality in a citizen science game: A design science evaluation. Proc Annu Hawaii Int Conf Syst Sci. 2013;450–9.
47. Good BM, Su AI. Crowdsourcing for bioinformatics. Bioinformatics. 2013; 29(16):1925–33. https://doi.org/10.1093/bioinformatics/btt333 PMID: 23782614
48. Ponciano L, Brasileiro F, Simpson R, Smith A. Volunteers' engagement in human computation for astronomy projects. Comput Sci Eng. 2014; 16(6):52–9.
49. Prestopnik N, Crowston K, Wang J. Gamers, citizen scientists, and data: Exploring participant contributions in two games with a purpose. Comput Human Behav. 2017; 68:254–68.
50. Sauermann H, Franzoni C. Crowd science user contribution patterns and their implications. Proc Natl Acad Sci. 2015; 112(3):679–84. https://doi.org/10.1073/pnas.1408907112 PMID: 25561529
51. Tinati R, Luczak-roesch M, Simperl E, Hall W. "Because Science is Awesome": Studying Participation in a Citizen Science Game. ACM Web Sci 2016. 2016;45–54.
52. Jiménez M, Triguero I, John R. Handling uncertainty in citizen science data: Towards an improved amateur-based large-scale classification. Inf Sci (Ny). 2019 Apr; 479:301–20.
53. Mavandadi S, Dimitrov S, Feng S, Yu F, Sikora U, Yaglidere O, et al. Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. PLoS One. 2012; 7(5):1–8.
54. Waldner F, Schucknecht A, Lesiv M, Gallego J, See L, Pérez-Hoyos A, et al. Conflation of expert and crowd reference data to validate global binary thematic maps. Remote Sens Environ. 2019; 221:235–46.
55. Surowiecki J. The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. London, UK: Little Brown; 2004.
56. See L, Comber A, Salk C, Fritz S, van der Velde M, Perger C, et al. DONT USE! Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. PLoS One. 2013; 8(7):1–11.
57. Luengo-Oroz MA, Arranz A, Frean J. Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. J Med Internet Res. 2012; 14(6):1–13.
58. Michelucci P. Science of Stall Catchers: Our new Magic Number [Internet]. EyesonAlz. 2017 [cited 2018 Nov 20]. Available from: https://blog.eyesonalz.com/our-new-magic-number/
59. Michelucci P. Validated Dynamic Consensus Approach for Citizen Science Projects Employing Crowd-based Detection Tasks. In: Presentation at the Citizen Science Association Conference 2017. Saint Paul, Minnesota; 2017.
60. Michelucci P, Dickinson JL. The power of crowds. Science (80-). 2016; 351(6268):32–3.

PAPER IV

Figure by Tara von Grebel

Strobl, B., S. Etter, H.J. van Meerveld, and J. Seibert (2020), Training citizen scientists through an online game developed for data quality control, *Geoscience Communication*, https://doi.org/10.5194/gc-3-109-2020.

Geosci. Commun., 3, 109–126, 2020
https://doi.org/10.5194/gc-3-109-2020

Research article

Training citizen scientists through an online game developed for data quality control

Barbara Strobl[1], **Simon Etter**[1], **H. J. Ilja van Meerveld**[1], **and Jan Seibert**[1,2]

[1]Department of Geography, University of Zurich, 8057 Zurich, Switzerland
[2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, 75007 Uppsala, Sweden

Correspondence: Barbara Strobl (barbara.strobl@geo.uzh.ch)

Received: 12 November 2019 – Discussion started: 2 December 2019
Revised: 11 March 2020 – Accepted: 26 March 2020 – Published: 8 May 2020

Abstract. Some form of training is often necessary for citizen science projects. While in some citizen science projects, it is possible to keep tasks simple so that training requirements are minimal, other projects include more challenging tasks and, thus, require more extensive training. Training can be a hurdle to joining a project, and therefore most citizen science projects prefer to keep training requirements low. However, training may be needed to ensure good data quality. In this study, we evaluated whether an online game that was originally developed for data quality control in a citizen science project can be used for training for that project. More specifically, we investigated whether the CrowdWater game can be used to train new participants on how to place the virtual staff gauge in the CrowdWater smartphone app for the collection of water level class data. Within this app, the task of placing a virtual staff gauge to start measurements at a new location has proven to be challenging; however, this is a crucial task for all subsequent measurements at this location. We analysed the performance of 52 participants in the placement of the virtual staff gauge before and after playing the online CrowdWater game as a form of training. After playing the game, the performance improved for most participants. This suggests that players learned project-related tasks intuitively by observing actual gauge placements by other citizen scientists in the game and thus acquired knowledge about how to best use the app instinctively. Interestingly, self-assessment was not a good proxy for the participants' performance or the performance increase through the training. These results demonstrate the value of an online game for training. These findings are useful for the development of training strategies for other citizen science projects because they indicate that gamified approaches might provide valuable alternative training methods, particularly when other information materials are not used extensively by citizen scientists.

1 Introduction

Citizen science projects can be grouped into two different types with regard to data collection and training: either citizen scientists are engaged in relatively straightforward tasks so that no training is needed, or they perform more advanced tasks that require detailed instructions and training (Breuer et al., 2015; Gaddis, 2018; Reges et al., 2016). Training needs depend on the tasks within the projects and the project organizers' perceived need for training. Environment-focused projects, in which citizen scientists perform simple tasks and, therefore, receive no prior training are, for example, the global project iNaturalist, where citizen scientists take a picture of plants and animals and upload it to a server (Gaddis, 2018; Pimm et al., 2014); CrowdHydrology, where people passing by a stream, such as hikers, read the water level of staff gauges in the USA (Lowry et al., 2019); a similar water level study in Kenya (Weeser et al., 2018); or a survey of the occurrence of hail in Switzerland (Barras et al., 2019). Projects in which citizen scientists receive training prior to being able to participate are, for example, CoCoRaHS (Community Collaborative Rain, Hail and Snow network), where citizen scientists operate a weather station (Reges et al., 2016); a groundwater study in Canada where volunteers measure the water level in wells (Little et al., 2016); a water quality study in Kenya and Germany (Breuer

et al., 2015; Rufino et al., 2018); or a water clarity study in lakes in the USA (Canfield et al., 2016). Bonney et al. (2009, p. 979) write "Projects demanding high skill levels from participants can be successfully developed, but they require significant participant training and support materials".

In practice, there is a range of citizen science projects, and many projects can be positioned between these two training types, especially when the tasks are relatively easy but data quality can be significantly improved with training. An example is Galaxy Zoo, which requires participants to classify galaxies in an online test, before they can start to submit data (Lintott et al., 2008). Another project is a malaria diagnosis game, which offers a short online tutorial for players (Mavandadi et al., 2012). Some projects offer in-person training (Kremen et al., 2011; Krennert et al., 2018; Rufino et al., 2018), but for many projects training has to be online because the projects are global (e.g., CrowdWater, Seibert et al., 2019a; CoCoRaHS, Reges et al., 2016; and an invasive-species training programme, Newman et al., 2010). Computer-based training can be tricky because the participants cannot be monitored. However, Starr et al. (2014) found that such computer training methods, e.g. via video, can be just as effective as in-person training. Computer-based training, furthermore, requires less time from the project organizers once the material has been developed.

The topic of training and learning in citizen science has received more interest in recent years (Bonney et al., 2016; Cronje et al., 2011; Jennett et al., 2016; Phillips et al., 2019). Many citizen science projects that provide training focus more on topic-specific knowledge often because this is required to complete the task successfully. Examples are the Flying Beauties project (Dem et al., 2018); the Neighbourhood Nestwatch programme (Evans et al., 2005); or invasive-species projects (Crall et al., 2013; Cronje et al., 2011; Jordan et al., 2011), where participants have to learn to identify species before they can participate in the project. However, some citizen science projects found that the participants did not increase their factual learning possibly because they were already quite advanced (Overdevest et al., 2004). Contributory projects often emphasize specific skills more than general topic knowledge. Examples of training for specific skills rather than knowledge are the Canadian groundwater study (Little et al., 2016) or the water quality study in Kenya (Rufino et al., 2018). However, "Engagement in contributory citizen science might, by way of the methods employed, result in more data reliability but fewer science literacy gains among participants." (Gaddis, 2018).

A novel approach to training was developed within the CrowdWater project. The CrowdWater project explores opportunities to collect hydrological data with citizen science approaches. On the one hand, the project develops new approaches to collect hydrological data by public participation (Kampf et al., 2018; Seibert et al., 2019a, b) and on the other hand assesses the potential value of such data for hydrological modelling (Etter et al., 2018; van Meerveld et al., 2017). In this study, the focus is on the collection of water level class observations based on the virtual staff gauge approach (Seibert et al., 2019a). This virtual staff gauge approach allows for water level observations without physical installations, such as staff gauges (Lowry et al., 2019; Weeser et al., 2018), so that it is scalable and can be used anywhere in the world. However, it is also more challenging for the user and potentially prone to mistakes (Seibert et al., 2019a; Strobl et al., 2019a). Previously we developed a web-based game for quality control of the water level class data (Strobl et al., 2019a). Here, we investigate whether playing this game might also be a useful preparation for using the virtual staff gauge approach in the CrowdWater app. The objective was to evaluate whether playing the game helped participants to understand the virtual staff gauge approach. More specifically, we addressed the following three questions:

- Are participants better at placing a virtual staff gauge after they have played the game?
- Are participants better at assessing the suitability of a reference picture after they have played the game?
- Are participants more confident in their contributions after playing the game, and is this confidence related to their performance in playing the game?

2 Background information on water level class observations in CrowdWater

2.1 CrowdWater app

The CrowdWater smartphone app enables citizen scientists to collect data for several hydrological parameters without requiring any physical installations or equipment. The app allows citizen scientists to set up new observation locations and to submit new observations for existing locations. The app uses OpenStreetMap (Goodchild, 2007) and thus allows for the georeferencing of observations worldwide. To start water level class observations at a new location, the citizen scientist takes a picture of a stream, showing the stream bank, a bridge pillar or any other structure that allows for the identification of the water level. Within the app, a virtual staff gauge is inserted onto this picture, which then becomes the reference for all further observations at this location (and is therefore called the reference picture). The virtual staff gauge is basically a sticker that is positioned as an additional layer onto the initial picture (Fig. 1a); i.e., there is no physical installation at the location. The citizen scientist can choose from three virtual staff gauges in the app, depending on the water level at the time when the picture is taken (low, medium or high; Seibert et al., 2019a). When placing the virtual staff gauge in the reference picture, the citizen scientist has to move the staff gauge so that it is level with the current water level and change the size of the staff gauge so that it covers the likely range of high and low water levels. When taking

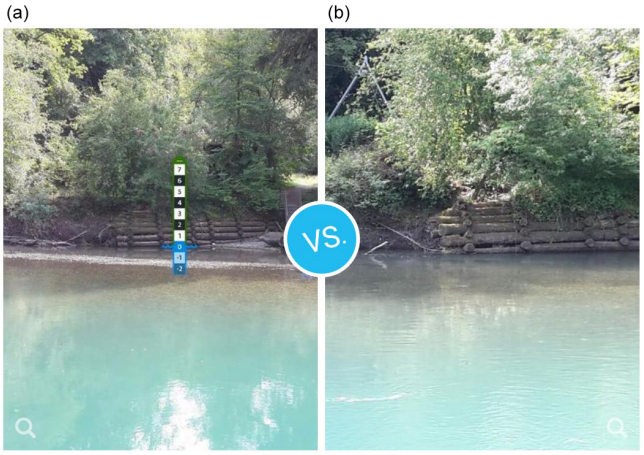


Figure 1. An example of a reference picture with the virtual staff gauge **(a)** and a picture from an observation at the same location at a later time **(b)**. The logs of the stream bank can be used as a reference to estimate the water level class.

the reference picture, it is important that it is perpendicular to the stream bank to avoid distortions when comparing the water level with the virtual staff gauge at a later time. Poor staff gauge placement is one of the most common errors and occurs in about 10 % of the new reference pictures (Seibert et al., 2019a). The most common errors are making the virtual staff gauge too big (or more rarely too small) to be useful to record water level fluctuations, not placing the staff gauge on the opposite river bank or perpendicular to the flow, or choosing the wrong staff gauge (Seibert et al., 2019a, b). For further observations (i.e., observations at the same location at a later time), the citizen scientist who created the reference picture or any other person who wants to report a water level class observation for this location looks for the structures in the reference picture (e.g., rock, bridge pillar or wall) and estimates the water level class by comparing it to the virtual staff gauge in the reference picture (Seibert et al., 2019a). In this way, time series of water level class data can be obtained at each observation location.

2.2 CrowdWater game

In addition to data collection using the CrowdWater smartphone app, citizen scientists can also contribute to the project by checking the collected water level class data in the web-based CrowdWater game (Strobl et al., 2019a). The idea of the CrowdWater game is to crowdsource the quality control of the submitted water level class observations by using the pictures that were taken and submitted by the citizen scientists in the app. In the game, picture pairs are shown: the reference picture with the virtual staff gauge and a picture of the same location at a later time (Fig. 1). The task is to estimate the water level class for the picture without the staff gauge (Fig. 1a) by comparing the water level in this picture with the reference picture, i.e. the picture with the staff gauge (Fig. 1b). Citizen scientists play rounds of 12 picture pairs: eight classified pictures that have already been assigned a "correct" value, i.e. the median based on the evaluations of at least 15 game players and four (so far) unclassified pictures. (This value is assumed to be the correct value but may diverge from the ground truth.) Currently, the CrowdWater game uses "unstructured crowdsourcing" (Silvertown et al., 2015, p. 127), which means that all votes are weighted equally to obtain the correct water level class. The order of the pictures is random so that the player does not know whether a picture pair has already been classified or not. For the classified picture pairs, points are obtained when the correct class (six points) or a neighbouring class (four points) is chosen, and zero points are given if the selected class is more than one class off from the correct value. For unclassified pictures, the player receives three points regardless of the vote. Players can also report a picture if voting is not possible because of, for instance, an unsuitable placement of the staff gauge, poor image quality or otherwise unsuitable pictures. In this case, the player also receives three points. The repeated evaluations of the same pictures by multiple players provide quality control of the incoming water level class data (Strobl et al., 2019a).

2.3 Motivation for this study

When using the CrowdWater app, citizen scientists take a picture of the observation location and upload it, similar to iNaturalist (Gaddis, 2018; Pimm et al., 2014) or iSpot (Silvertown et al., 2015). When starting observations at a new location, some interpretation is needed, which requires an understanding of the possible range of water levels and determination of the current water level. The data collection protocol is, however, simpler than for many projects that do require training; therefore low-intensity training seems to be advisable for the CrowdWater project.

As a first step, manuals (https://www.crowdwater.ch/en/crowdwaterapp-en/, last access: 30 April 2020) and instruction videos (https://www.youtube.com/channel/UC088v9paXZyJ9TcRFh7oNYg, last access: 30 April 2020) were provided online, but in our experience (and based on the number of views on YouTube) these are not frequently used. Thus some citizen scientists occasionally still make mistakes when submitting data in the CrowdWater app, primarily when starting a new location for observations and placing a virtual staff gauge onto the reference picture (Seibert et al., 2019a). Our first approach to handle these mistakes was to implement a method of quality control to either filter out or correct erroneous submissions. This quality control method was gamified in the CrowdWater game. The CrowdWater game proved successful in improving the quality of the water level class data submitted through the app (Strobl et al., 2019a). Shortly after launching the game, we received anecdotal evidence, such as direct feedback

from players, that the game also helped them to better place staff gauges and to better estimate water level classes. This feedback was confirmed through a short survey sent out to CrowdWater game players for a different study (Strobl et al., 2019a). Roughly a quarter of all players at the time filled in the survey (36 players). When asked if playing the game helped them to be more aware of how to place a staff gauge in the app, 79 % agreed. Furthermore, 58 % of all surveyed players agreed that the game helped them to better estimate water level classes in the app. The other players indicated no change in their abilities, and none of the survey players indicated a deterioration of their skills. Essentially, the players are training each other in the game as the score per picture pair, which is based on the votes of the other players, shows the new player if they are correct or not. This is similar to iSpot, where experts train beginners in species recognition (Silvertown et al., 2015). Through the CrowdWater game, players learn which staff gauges are difficult to read and which ones allow for an easy comparison of the water levels (Strobl et al., 2019a).

This motivated us to investigate if the CrowdWater game can be used to train potential citizen scientists to place the virtual staff gauge in the CrowdWater app correctly. It is better to train citizen scientists before participation so that they provide useful data rather than to filter data from untrained citizen scientists afterwards. Filtering wrong data afterwards wastes the time of the citizen scientists, and erroneous data can be missed by the filter. In the CrowdWater project, it is particularly important to place the virtual staff gauge correctly because all subsequent observations at an observation location are based on this virtual staff gauge (i.e., a poorly placed staff gauge will influence all following observations).

The CrowdWater game is a project-specific training tool meant to improve the reliability of CrowdWater observations and does not aim to improve scientific literacy. This is similar to some other citizen science projects, especially contributory projects, where data are crowdsourced (Crall et al., 2013). Improving the hydrological knowledge was not necessary in our case, as the data can easily be collected without such background knowledge. However, other materials that provide such knowledge and a link to an open massive online course are provided on the project website.

3 Methods

3.1 Training study

This study aimed to assess if the CrowdWater game can be used to train new participants to place the virtual staff gauge in the CrowdWater app correctly. The placement of the staff gauge is the most important metric for this study because this is the most crucial task when CrowdWater app users start a new observation location. Rating reference pictures gave additional insight into whether participants can recognize well and poorly placed staff gauges, regardless of whether or not they can place them well themselves.

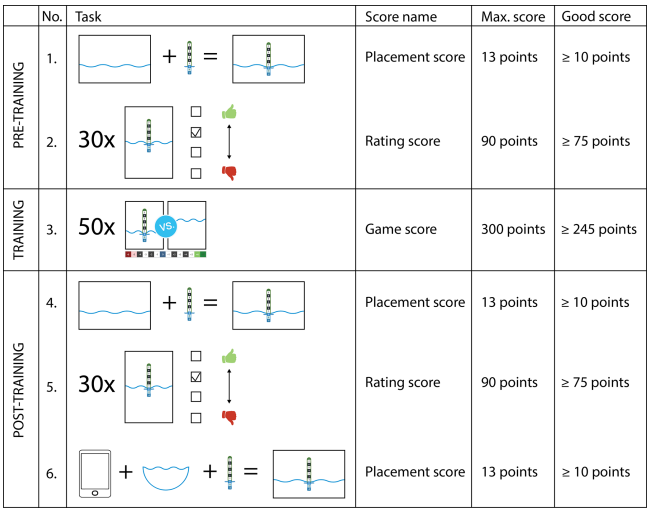


Figure 2. Schematic overview of the pre-training, training and post-training tasks. For each task, the maximum number of points and the chosen value for good performance are given.

The training study consisted of a number of tasks that were executed before and after playing the game. To focus on the research questions and to exclude other factors, such as differences between locations, flow conditions or daylight, the study was mainly conducted indoors at a computer. For each participant, the experiment took 60–90 min. All instructions and questions were formulated in English; all participants had a good command of English. The study was conducted between August and October 2018, apart from a small outdoor task, which was completed by the participants at a later time. The full study can be found in the "Training study" in the Supplement.

3.1.1 Study tasks

The six tasks of the training study can be divided into pre-training, training and post-training tasks (Fig. 2). Each participant completed these tasks in the same order. Pre-training and post-training tasks are only intended to assess the participant's performance during this study and are not part of the training for the CrowdWater project.

The pre-training tasks are structured as follows:

– *First task (staff gauge placement)*. The study participant looked at 18 stream pictures of the river Glatt (see the stream pictures in the Supplement). The pictures show the same location but were taken from different angles and perspectives. Some were well suited for placing a virtual staff gauge; others were moderately suitable; and some were not suitable at all. Without receiving any further information, the participant was asked to choose

1 of the 18 pictures and to place a virtual staff gauge onto the picture. This was done using an interface on the computer that looked similar to that in the CrowdWater app.

– *Second task (rating of reference pictures)*. The participant looked at 30 different reference pictures (for examples see "Examples of reference pictures for the rating task" in the Supplement). These pictures were chosen from reference pictures that were uploaded by citizen scientists using the CrowdWater app. The pictures were selected to represent a range of well, moderately and poorly placed virtual staff gauges. The participant rated each of the 30 reference pictures as "unsuitable", "rather unsuitable", "rather suitable" or "suitable".

The training task is structured as follows:

– *Third task (game)*. The participant played an adapted version of the CrowdWater game. In this version, the participant estimated the water level class of 50 picture pairs. The regular CrowdWater game only offers 12 picture pairs per day, so this extended version corresponds to the training effect of about four rounds of the game. The participant did not receive any explanation on the game but could use the help button to obtain more information on the game.

The post-training tasks are structured as follows:

– *Fourth task (staff gauge placement)*. The participant repeated the first task and was asked to place the virtual staff gauge for the river Glatt again. The participant received the same 18 pictures but was free to choose another picture and to place the virtual staff gauge in a different location, angle or size compared to the first task or to choose the same picture and to place the staff gauge similarly.

– *Fifth task (rating of reference pictures)*. The participant repeated the second task for a different set of 30 reference pictures from the app. The distribution of well, moderately and poorly placed virtual staff gauges was roughly the same as in the second task.

– *Sixth task (staff gauge placement)*. The participant used the CrowdWater app outdoors (instead of the online interface used for the earlier tasks) to create and upload a reference picture for a stream of their choice. The task was meant to be completed within 2 weeks after completing the first five tasks. However, not every participant completed the task within this timeframe (at the latest by March 2019), and 10 participants did not complete this task at all.

After placing the staff gauge online (first and fourth task) and rating the reference pictures (second and fifth task), participants answered several questions to assess the difficulty of the task, their own performance, and their confidence in completing these tasks correctly. After the training (third task), participants were asked about the difficulty of the game and whether they thought the game was fun.

3.1.2 Assessment of the different tasks

The performance of the participants for the different tasks was evaluated based on a score. The scores before and after playing the game (i.e., the training) were compared to determine the effect of playing the game. The scoring system was determined prior to the start of the study according to assessment criteria that were based on previous experiences with pictures submitted through the app and expert judgement (by Barbara Strobl and Simon Etter). A separation of the individual scores into "good" and "poor" was, while somewhat arbitrary, necessary to be able to distinguish the effects of the training on the participants who needed it most, i.e. those who had poor performance (i.e., score) before the training.

For the staff gauge placement tasks (first, fourth and sixth task), points were given for five different placement criteria. The maximum placement score was 13. A placement score of 10 or higher was considered good because these reference pictures can still be used and would have been left in the CrowdWater database if they were submitted through the app (Fig. 3).

Perspective of the picture. The 18 pictures of the river Glatt were taken from different angles and perspectives and assigned a score: 0 (unsuitable), 1 (rather unsuitable), 2 (rather suitable) and 3 (suitable). The participant could gain more points for the choice of the picture than the other criteria for placing a staff gauge because this is essential for a good reference picture. Because every participant fulfilled the outdoor task (sixth task) for a different stream and the participants could choose a location themselves, points could not be assigned a priori. However, the location and the picture frame were assessed, and a score between 0 and 3 was given based on expert judgement (by Barbara Strobl and Simon Etter).

Choice of the staff gauge. Participants could choose from three different virtual staff gauges depending on the water level at the time that the picture was taken (low, medium or high). The staff gauge for low flow was considered correct, as the water level was low at the time that the 18 pictures of the Glatt were taken. The score for the selected staff gauge varied between two (staff gauge for low flow), one (staff gauge for medium flow) and zero (staff gauge for high flow). For the outdoor task with the app (sixth task), the situation was assessed based on the water level, and points were assigned for the correct assessment of low, medium or high flow by the participant.

Location of the staff gauge. If the staff gauge was placed on the opposite stream bank, as it should be, two points were given; if the staff gauge was incorrectly placed on the participant's side of the stream or in the middle of the stream, zero points were given.

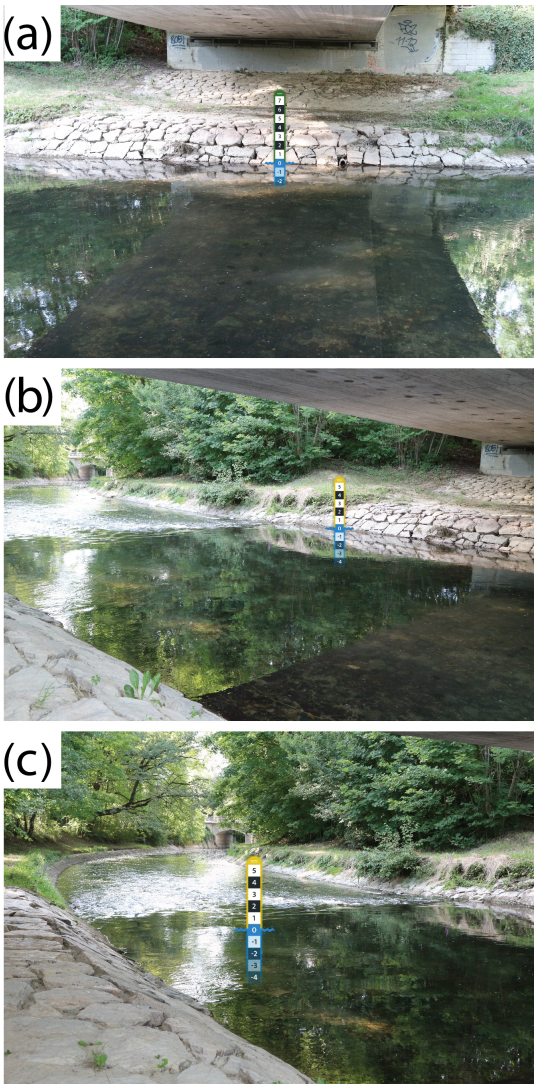


Figure 3. Examples of staff gauge placements: **(a)** 13 points, i.e., a full score; **(b)** 10 points, just enough points to still be considered suitable for future water level observations; and **(c)** 3 points, the lowest score obtained throughout the study.

Angle of the staff gauge. The staff gauge should be placed perpendicular to the flow in the stream to avoid contortions of the perspective for future water level class estimates. If the staff gauge was placed perpendicular to the flow ($\pm 10°$), two points were given. If the angle was less than 45°, one point was given, and if it was larger than 45°, zero points were given.

Water level mark. The blue wave of the staff gauge should be located at the water surface in the reference picture. If this was the case, two points were given; if the blue wave was only slightly off, e.g., due to reflections on the water surface, one point was given; and if the blue wave was not placed on the water surface, zero points were given.

Very rarely, two virtual staff gauges were placed in the reference picture (twice before the training – first task, once after the training online – fourth task – and once after the training outdoors in the app – sixth task). We assume that this was most likely due to technical difficulties. In these cases, we subtracted one point from the participant's score. This, however, had hardly any effect on the results.

The rating of the reference pictures (second and fifth task) was evaluated using a rating score. The participant's choice between unsuitable, rather unsuitable, rather suitable and suitable was compared to the expert judgement of the reference pictures (by Barbara Strobl and Simon Etter). If the participant picked the same suitability class as the experts, three points were given. For each class deviation from the expert judgement, one point was subtracted. Thus the maximum score was 90 points (30 reference pictures times 3 points per picture). A score of 75, which corresponds to being off by one class 5 times and off by two classes 5 times and choosing the correct class 20 times, was still considered good.

For the training task (fifth task, the game), the participants received points for each picture pair that they compared. Similar to the actual CrowdWater game, they received six points if they chose the correct class, i.e. the median of the votes of all previous CrowdWater game players; four points if they chose a water level class that was one class away; and zero points if they chose a class that was more than one class away from the median. When reporting a picture pair, the participant received three points. The maximum score for the training task was 300 points (a maximum of 6 points times 50 picture pairs). The threshold for a good game score was determined before the study and set at 245 points, which reflects a situation where a participant chose the correct class for 35 out of the 50 picture pairs, was one class off five times, was more than one class off for another 5 picture pairs and reported five pictures (we considered five pictures unsuitable and would thus have reported them).

3.1.3 Data analysis

The scores for the staff gauge placement and rating tasks before and after the training were compared for each participant using two paired statistical tests: the paired sample t test for normally distributed data and the Wilcoxon test for data that were not normally distributed (Table 1). We used a one-sided test to check whether the difference in the scores before and after the training was larger than zero and a two-sided test to determine the significance of the difference in the scores between the computer-based and the outdoors app-based staff gauge placement (i.e. between the fourth and the sixth task). We used a significance level of 0.05 for all tests. We performed the tests for all participants together but also divided the participants based on their placement score before the training (first task) in order to determine the effect of training for people who initially did not install the virtual

staff gauge correctly. In order to see whether the game performance was related to the improvement in the placement or rating score, we also split the data based on the game score. We used Spearman rank correlation (r_s) to evaluate the relation between performance (i.e., scores) and the confidence of the participants in their performance, as well as between the performance and the stated difficulty and fun rating.

3.2 Study participants

The participants for this study were recruited through various channels. The University of Zurich offers a database with potential study participants in the vicinity of Zurich; people in this database were contacted via email. Additional emails were sent to staff and students of the Department of Geography. Friends, colleagues and family helped to recruit participants from their social network as well. Local study participants could complete the online part of the study in a computer room at the University of Zurich at specified times; all other participants received the link and completed the study on their own. All participants completed the first five tasks individually in one session.

The participants in this study had neither previously used the CrowdWater app nor played the CrowdWater game. In total, 52 participants completed the first five tasks of the study. Of these 52 participants, 10 did not complete the outdoor app task, but their results were included in the analyses as far as possible. When sending email reminders to complete this sixth task, several participants indicated a lack of time or a suitable nearby river. Most of the 10 participants intended to complete the task but forgot about it in the end. Of the 52 participants, 32 (62 %) were female, and 20 (38 %) were male. Age data were collected in age groups: 6 % of the participants were under 20 years old; 79 % of the participants were 21–40 years old; 8 % were 41–60 years old; and 8 % were 61–80 years old. The highest education level was secondary school for 4 % of the participants, high school for 12 % of the participants, university (BSc, MSc or similar) for 79 % of the participants and a PhD for 6 % of the participants. The education level being higher than the Swiss average and the relatively large group of young people (< 40 years) are due to the recruitment of the participants at the University of Zurich. The education level of the CrowdWater citizen scientists is unknown, but 89 % of the 36 CrowdWater game players who filled in a survey about the game were university educated, and 75 % were under the age of 40 (Strobl et al., 2019a). For a survey about the motivations of CrowdWater app users, as well as citizen scientists from a different phenological citizen science project (Nature's Calendar ZAMG; Zentralanstalt für Meteorologie und Geodynamik), 66 % of the respondents were university educated, and 51 % were under the age of 40 (Etter et al., 2020).

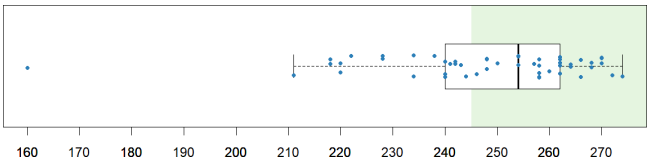


Figure 4. Boxplot of the game score for each study participant. Scores ≥ 245 points are considered good (indicated by the green background). The box represents the 25th and 75th percentile; the line is the median; the whiskers extend to 1.5 times the interquartile range. The individual scores (blue dots) are jittered to improve the visibility of all points.

4 Results

4.1 Training results

Almost two thirds (62 %) of the study participants had a good score (≥ 245 points) for the game. The highest game score was 274, and the average score was 248. The lowest score (160 points) was an outlier; the second-lowest score was 211 points (Fig. 4). Interestingly the participant with the lowest game score found the game "rather difficult" but still "a bit of fun", adding "It [the game] was quite tricky. I was curious if my answer is right or wrong".

In the game, participants can report a picture pair if they think that it is not possible to vote on a water level class. The reason for reporting a picture pair can be selected from a drop-down menu. The report function was used by 16 participants (31 %). It is unknown if the other 36 participants did not find the report function or if they did not think it was necessary to report any of the picture pairs. Most of the participants who used the report function reported between 1 and 6 picture pairs, but one participant reported 10, and another participant reported 12 picture pairs. Out of the 50 picture pairs in the game, 22 were reported at least once, and 1 picture pair was reported seven times. When choosing the 50 picture pairs for the game, we included 5 picture pairs that should be reported (Fig. 5). In other words, there were 57 reports in total, 38 of which were not valid (i.e., our expert knowledge suggests that the picture pairs could be used to determine the water level class). For some of these cases, participants considered a spot unsuitable because they did not realize that they could see the entire picture if they clicked on it and therefore thought the reference picture did not have a staff gauge. In another case, they may have been confused by a slightly different angle in the picture for the new observation. The most common reason for reporting a picture was "The location has changed, and the reference image is unrecognizable". This was indeed a problem with some of the picture pairs (Fig. 5).

Table 1. The statistical tests were chosen based on whether or not the data were normally distributed according to the Shapiro–Wilk test. The tests for the placement score compared scores from before and after the training, as well as after the training and outdoors with the app. The test for the rating score compared scores from before and after training.

Data	Data subset	Results of the Shapiro–Wilk test	Statistical test of the training effect
Placement score	All participants	Not normally distributed	Wilcoxon test
	Participants with a low placement score before the training	Not normally distributed	Wilcoxon test
	Participants with a good game score	Not normally distributed	Wilcoxon test
	Participants with a bad game score	Not normally distributed	Wilcoxon test
Rating score	All participants	Normally distributed	Paired sample t test
	Participants with a low rating score before the training	Not normally distributed	Wilcoxon test
	Participants with a good game score	Not normally distributed	Wilcoxon test
	Participants with a low game score	Normally distributed	Paired sample t test

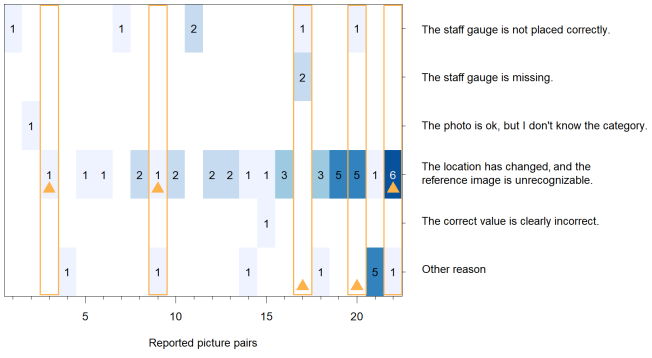


Figure 5. The number of times that a picture pair was reported and the reason for reporting the picture pair (y axis) for the 22 picture pairs in the game that were reported at least once (x axis). The picture pairs that should have been reported based on expert assessment prior to the training study are framed with an orange rectangle; the orange triangle indicates the reason based on expert assessment. The blue shading represents the number of reports per picture pair (as also indicated by the printed number).

4.2 Staff gauge placement

4.2.1 Placement scores before training

The staff gauge placement score before the training (first task) was 10 or higher for 70 % of the participants; i.e., the majority of the participants placed the staff gauge in a way that is suitable for further observations. This is a good performance considering that the participants had not yet received any training. Training is more important for the 30 % of participants who had a low placement score before the training. The lowest scores were two points (one participant) and three points (two participants).

4.2.2 Placement scores after training

The placement scores generally improved after the training and were statistically significantly better than the scores before the training (Wilcoxon test, $p<0.01$; Fig. 7). Improvement is especially important for the participants who had a low placement score before the training. Therefore, the participants with a low initial score (<10 points) were assessed separately. For this group, the median placement score improved significantly with training as well (Wilcoxon test, $p<0.01$). Of the 16 participants with a poor placement score before the training, 10 improved their staff gauge placement sufficiently to make it useful for future observations. Participants who performed well before the training have less possibility of improving the placement and also need to improve their placement score less. However, for two of the participants with a good score before the training, the score was poor after the training (Figs. 6 and 7). The placement score improved for participants with a good game score (Wilcoxon test, $p<0.01$) but not for participants with a low game score (Wilcoxon test, $p=0.11$).

A total of 19 participants (37 %) picked a different picture for the staff gauge placement after the training. Eight of these participants chose a stream picture with the same suitability

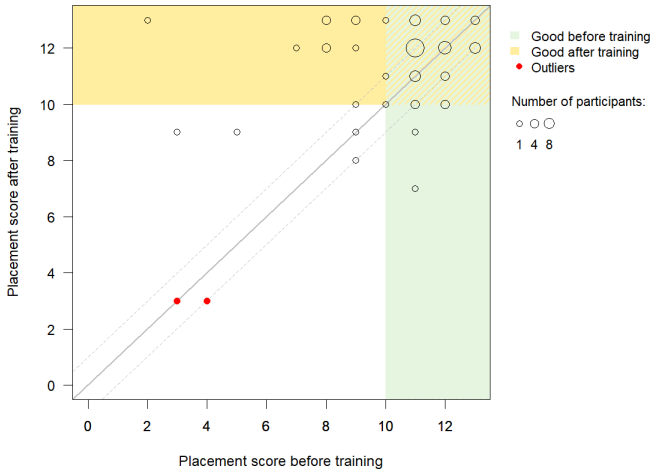


Figure 6. Placement scores before the training (x axis) and after the training (y axis). The circle size indicates the number of participants with the same scores. The green background indicates participants who had already performed well in placing the staff gauge before the training (score ≥ 10), and the yellow background indicates participants who performed well after the training (score ≥ 10). The solid grey line indicates the 1 : 1 line (i.e., the same score before and after the training), while the dashed lines indicate a difference of only one point. Points in the upper-left triangle indicate an improvement in staff gauge placement after the training. The red circles indicate outliers.

score as the first one; nine selected a better stream picture; and two chose a picture that was worse than their original choice. The other 33 participants chose the same stream picture as before. The good scores even before the training suggest that most of them also did not need to change the picture. The participants who changed the stream picture had a median placement score of 9 before the training and 12 after the training. The participants who chose the same stream picture had a median placement score of 11 before the training and 12 after the training. Before the training, 37 participants chose a reference picture with a score of 3, 9 with a score of 2, 5 with a score of 1, and only 1 participant chose a reference picture with a score of 0. Of the six participants who had a score of 0 or 1 before the training, four chose a reference picture with a score of 2 or 3 after the training. For two participants the reference picture score remained 1.

Except for one participant, all participants who performed well in the training task (game score ≥ 245 points) had a good placement score (≥ 10) after the training. However, the opposite was not the case: participants with a low game score (<225 points) sometimes still improved their placement score after the training, and all had a good placement score (≥ 10) after the training (Fig. 8). The participant with the most substantial improvement in staff gauge placement (from 2 points to 13 points) had an excellent game score of 262 points (Fig. 8). Participants who obtained a low score for the staff gauge placement after the training all had an average

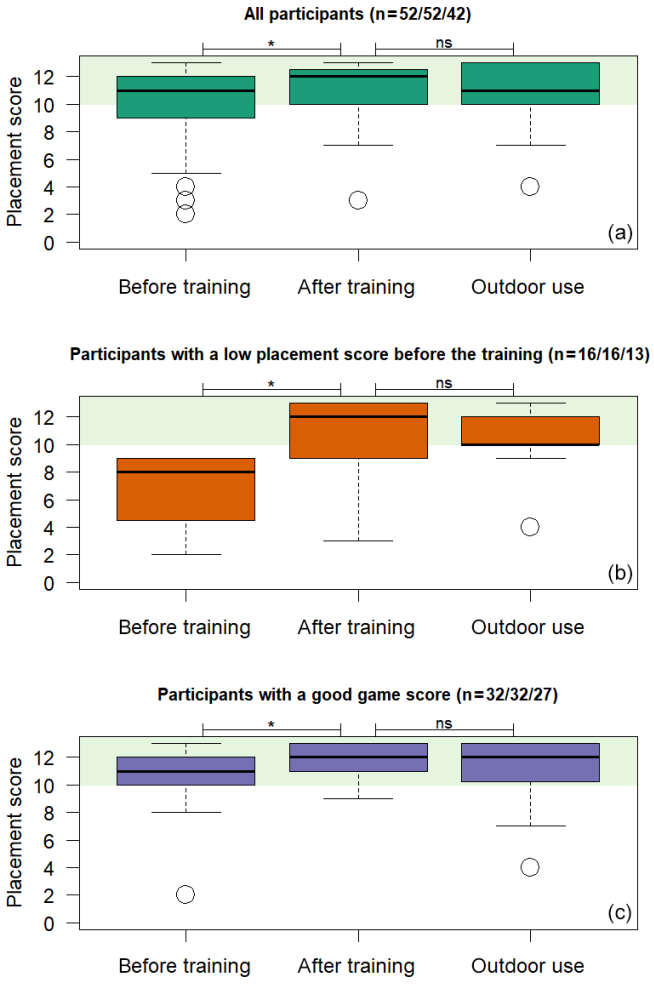


Figure 7. Box plots of the placement scores before the training (first task), after the training online (fourth task) and outdoors with the app (sixth task) for all participants **(a)**, for participants who had a low placement score before the training **(b)** and for participants who had a good game score **(c)**. There was a statistically significant difference in the placement scores before and after the training for all groups (indicated by *) and no statistically significant difference between the computer-based task and the outdoor app task (indicated by "ns") after the training based on the Wilcoxon test ($p<0.05$). The green shading indicates a good score.

score in the game (228–243), except for one participant with a high game score (248; Fig. 8).

There was no statistically significant difference between placement scores after the training for the online (fourth task) and the outdoor task with the app (sixth task), neither for all participants ($p = 0.50$), for participants with a low placement score before the training ($p = 1.00$), nor for participants with a good game score ($p = 0.20$) or for participants with a bad game score ($p = 0.57$; Fig. 7). This indicates that the online task can be used as a proxy for handling the app.

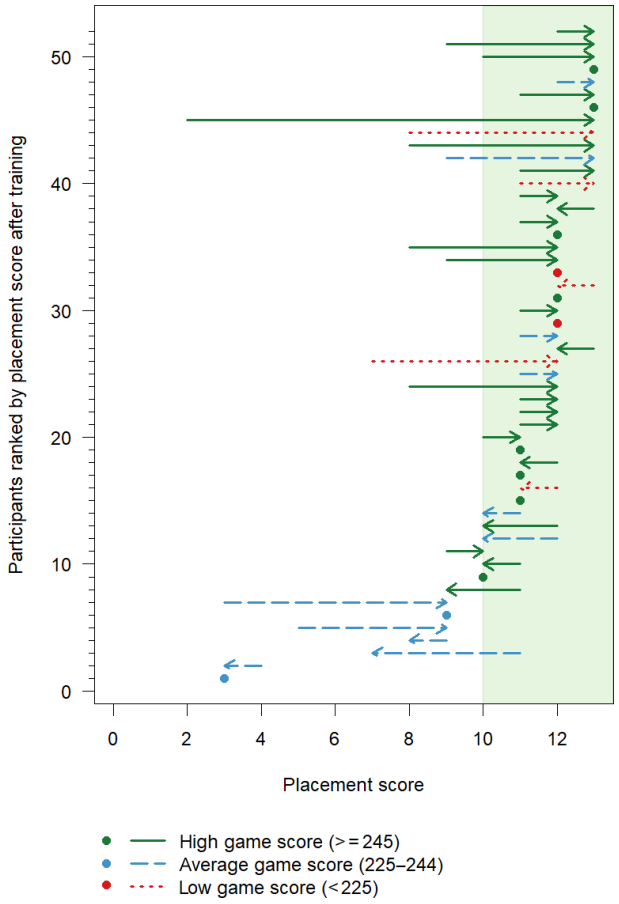


Figure 8. Placement scores before and after the training (x axis) per participant (y axis). Arrows point to scores from before to after training. Dots indicate no change in the placement score and are coloured according to the game score they obtained during the training.

4.2.3 Placement score outliers

When plotting the placement score before the training and after the training, two outliers were visually identified (Fig. 8). Both participants had a low score before the training and, unlike other participants, also a low score after the training. These two participants received few points across all assessment criteria for staff gauge placement and also had a below-average game score (242 and 228 points). They rated the game as rather difficult and very difficult, and when asked whether they enjoyed playing the game, they rated it neutral and stated that "It wasn't fun at all". Surprisingly both participants were confident that the reference picture for the staff gauge placement was rather suitable. Both participants changed their impression of the difficulty of the staff gauge placement (first task) from very easy before the training to rather easy and neutral after the training (fourth task).

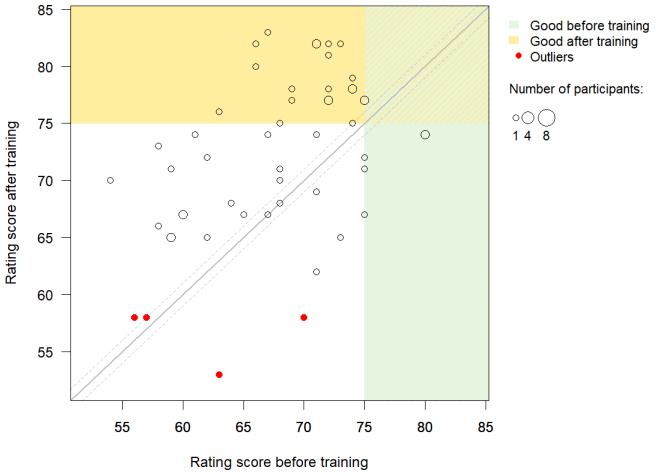


Figure 9. Rating scores before the training (x axis) and after the training (y axis). The circle size indicates the number of participants with the same scores. The green background indicates participants who had already performed well (score ≥ 75) before the training, and the yellow background indicates participants who performed well after the training (score ≥ 75). The solid grey line indicates the 1 : 1 line (i.e., the same score before and after the training), while the dashed lines indicate a difference of only one point. Points in the upper-left triangle indicate an improvement in the rating score after the training. The red circles indicate outliers.

4.3 Rating of reference pictures

4.3.1 Rating scores before the training

Even though the majority of the participants received a good staff gauge placement score before the training, only 13 % of the participants had a good rating score (≥ 75) before the training. The highest rating score before the training was 80, and the lowest score was 54; the average score was 68 points. Only 9 % of the participants had a good score for both staff gauge placement and rating before the training.

4.3.2 Rating scores after the training

The rating scores improved after the training (Figs. 9 and 10). The median difference in the rating score before and after the training was statistically significantly larger than zero, for all participants (paired-sample t test, $p<0.001$), for participants with a low rating score before the training (Wilcoxon test, $p<0.001$), for participants with a good game score (Wilcoxon test, $p<0.001$) and for participants with a low game score (paired-sample t test, $p = 0.02$; Fig. 10).

The rating scores can also be analysed per picture. A single picture can receive between 156 points (all 52 participants chose the correct suitability class and received 3 points) and 0 points (all participants chose the suitability class that is furthest from the correct class). The score was higher for the reference pictures that were considered to be unsuitable by the experts before the study (median: 139; range: 77–152) than

for the pictures that the experts rated as suitable, rather suitable and rather unsuitable (median: 120–121). This indicates that participants were better at identifying the unsuitable pictures than the more suitable pictures (Table 2).

4.3.3 Rating score outliers

Outliers for the rating scores were less obvious than for the placement scores, although there appear to be four outliers (Fig. 9, red circles). One participant was also an outlier for the staff gauge placement. The game scores and the assessment of difficulty and fun of the game varied for these four participants. The confidence in their own performance when rating the reference pictures was mixed before the training, but it was never lower than neutral. After the training, all four participants were confident in their performance and found the task either rather easy or very easy.

4.4 Confidence, difficulty and fun

4.4.1 Confidence and difficulty in staff gauge placement and rating the reference pictures

The participants were in general quite confident in their performance, and their confidence increased after the training (from 67 % to 98 % of participants for staff gauge placement and from 62 % to 90 % for rating the reference pictures; Fig. 11). As shown above for the outliers in the placement score and rating score, the participants' confidence in their performance was not correlated with their actual performance, neither before nor after the training ($|r_s| \leq 0.23$; $p \geq 0.11$).

Before the training, participants thought that the placement of the staff gauge was a relatively easy task, but the level of difficulty was roughly equally split between difficult, neutral and easy for the rating of the reference pictures (Fig. 11c, d). Participants generally considered the tasks easier after the training (72 % of the participants said that the placement of the staff gauge was easy before the training vs. 84 % of the participants after the training; 43 % of the participants thought that rating the reference pictures was easy before the training vs. 71 % after the training). Similar to the results for confidence, the assessment of the difficulty of the task was not related to the performance, neither before nor after the training ($|r_s| \leq 0.16$; $p \geq 0.30$).

4.4.2 Difficulty and fun of the game

Of the participants, two thirds thought that playing the game was fun, but when rating the difficulty, they were almost equally split between difficult, neutral and easy (Fig. 12). All participants who thought that the game was not fun (21 %) thought that the game was either difficult or neutral. The level of fun and difficulty was correlated ($r_s = 0.43$; $p < 0.01$). Nonetheless, 11 % of the participants stated that they had fun during the game but also thought it was difficult.

4.5 Feedback

Participants had the option to provide unstructured feedback at the end of the online study (after the fifth task); 15 participants decided to do so. Five participants mentioned different issues that had been unclear to them during the study, and four commented that they had enjoyed taking part in the study; two specifically mentioned that they thought that the training had helped them to understand the virtual staff gauge approach, but one participant stated that they thought the training had not helped. Two participants stated that they thought the study was difficult, and two gave feedback on the technical implementation of the study.

5 Discussion

5.1 Does the CrowdWater game help participants to place the virtual staff gauge in a suitable way?

The virtual staff gauge approach was developed as an intuitive approach to collect water level data so that many citizen scientists would be able to contribute observations to the CrowdWater project. Such a simple approach is often recommended to citizen science project initiators (Aceves-Bueno et al., 2017). Many other citizen science projects, such as CrowdHydrology and iNaturalist, also deliberately chose to keep the data collection method easy so that citizen scientists do not require training prior to participation (Gaddis, 2018; Lowry et al., 2019).

When starting a new CrowdWater location for water level class observations, the most difficult task is placing the staff gauge. This is also the first thing that most citizen scientists who use the CrowdWater app do. Recording follow-up observations in the app is much easier than placing the virtual staff gauge. However, the staff gauge placement is an essential task, as all subsequent observations of water level classes are based on the reference picture. A well-placed staff gauge makes the subsequent observations easier, more reliable and more informative. This is not ideal, as the citizen scientist might not have fully understood the concept of the virtual staff gauge yet when making the first observation. Mistakes in the placement of the virtual staff gauge occur in about 10 % of the cases.

In this study, most participants (70 %) were already good at placing a staff gauge, even before receiving any training. This indicates that the virtual staff gauge is indeed intuitive to use. Training is especially important for the participants who did not place the staff gauge well before the training, i.e., citizen scientists who do not intuitively understand how to place the staff gauge in the app. Starr et al. (2014) reached a similar conclusion in a study that compared different training methods for plant identification and also focused on the beginner group to see the training effects clearly. While the CrowdWater app is reasonably intuitive, the fact that we do sometimes receive submissions with mistakes (Seibert et al.,

Table 2. Number of pictures to be rated before and after the training per suitability category (as determined prior to the study by the experts) and the median, average and range in rating scores for the pictures in each category. Each picture can receive a maximum rating of 156 points (i.e., all 52 participants chose the correct category and therefore gained three points).

	Number of pictures		Rating score (0–156)		
Suitability category	Before training (second task)	After training (fifth task)	Median	Average	Range
Unsuitable	8	8	139	133	77–152
Rather unsuitable	4	3	121	118	105–128
Rather suitable	6	6	121	119	96–132
Suitable	12	8	120	116	66–138

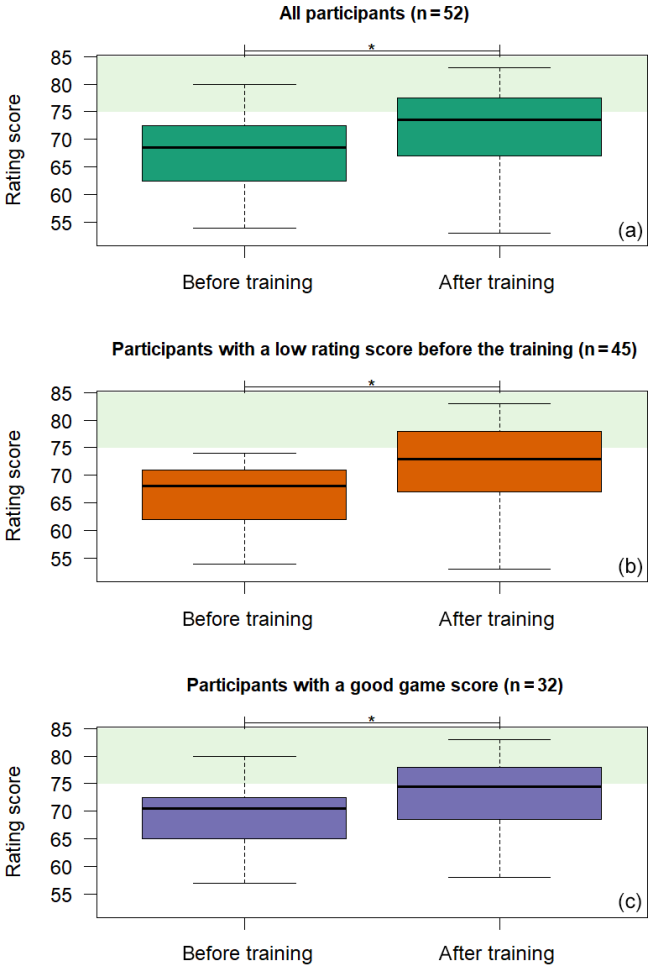


Figure 10. Boxplots of the rating score before and after the training for all participants **(a)**, for participants who had a low rating score before the training **(b)** and for participants who had a good game score **(c)**. The difference was statistically significant for all groups based on the Wilcoxon test ($p<0.05$; indicated by *). The green shading indicates a good rating score.

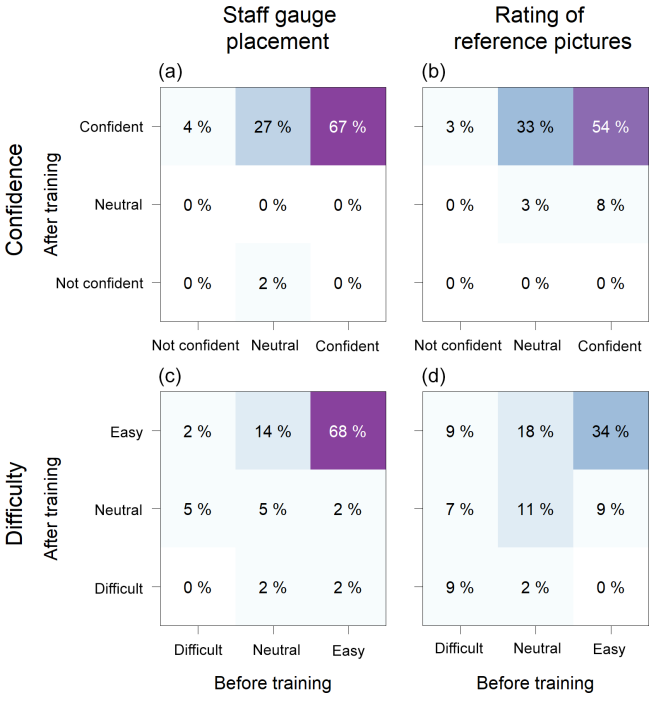


Figure 11. Percentage of participants who chose a certain confidence level **(a, b)** and their assessment of the difficulty of the task **(c, d)** for the staff gauge placement **(a, c)** and rating of reference pictures **(b, d)** before the training (x axis) and after the training (y axis). Darker colours indicate that a higher percentage of participants chose these options.

2019a) suggests that training could be beneficial. The mistakes made when using the app, closely resemble the mistakes made by participants in this study and included making the staff gauge too big, not placing the zero line on the water level, or choosing a picture with an angle that distorts the image and hampers further observations at this location. Playing the CrowdWater game can help to avoid these mistakes in a playful manner for some of the participants (63 % of the participants who performed poorly prior to training did well after training). Based on these findings, we suggest that new

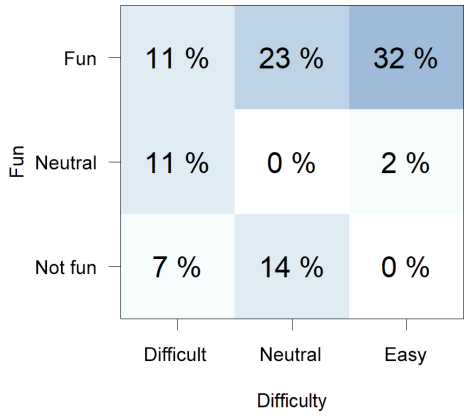


Figure 12. Percentage of participants who chose a certain category for the difficulty (x axis) and fun (y axis) of the game. Darker colours indicate that more participants chose these options.

citizen scientists play the CrowdWater game before setting up a new observation location.

Playing the CrowdWater game was not helpful for all participants; some participants who had a low placement score before the training had a low placement score after the training as well. Rinderer et al. (2015) reported a similar case, where some groups did improve their skills at classifying soil moisture, but others did not. In the context of this study, this might be due to the CrowdWater game being an implicit approach to training, instead of an explicit one. We did not provide theory about staff gauge placement nor mention the essential criteria of a good virtual staff gauge placement to participants (e.g. angle, size or placement on water level) during the study. Most participants intuitively understood this after playing the game because they noticed that a poor placement of the staff gauge made the estimation of the water level classes for subsequent observations more difficult. The benefit of such an implicit approach is that it is likely more fun than merely providing the theory (which is given on the CrowdWater website and explained in instruction videos). Nonetheless, some participants might have preferred explicit, written instructions on what to look for, instead of having to acquire this knowledge themselves. We, therefore, recommend that citizen science projects offer theoretical material in addition to a gamified training approach. Newman et al. (2010) encourage citizen science project leaders to provide many different training approaches to accommodate different learning styles. We do not know if the participants who benefited most from playing the game had previous experiences with citizen science, online games or smartphones. This could be investigated in a future study and would indicate who might require more training or for whom training via a game is most beneficial.

When rating the reference pictures, participants were better at recognizing unsuitable reference pictures compared to rather unsuitable, rather suitable or suitable pictures. The boundaries between the intermediate categories are of course vague and somewhat subjective, but it is very encouraging that participants could accurately identify unsuitable reference pictures, as this means that they are aware of what constitutes a poor placement and are therefore less likely to make these mistakes themselves. This is slightly contradictory to the results on the use of the report function during the game. While few participants reported pictures, those who did often overused this opportunity and reported more picture pairs than needed. In practice, it is tricky to decide where to set the limit between a suitable and unsuitable picture. For the majority of the reference pictures submitted via the CrowdWater app, the staff gauge placement is neither perfect nor useless. Although many staff gauges are not placed ideally, this does not necessarily mean that they are unusable. Depending on the location, it is often also not possible to place the virtual staff gauge perfectly.

There was no strong correlation between the game score and the improvement after the training. This could partly be due to the fact that learning occurs gradually during the game. Early in the game, participants might get few points and improve later during the game, leading to an average game score and a learning effect before finishing the training. The number of game rounds for optimal training is unknown, but the four rounds used here may be a good compromise between showing enough different pictures and not taking too much time. Strobl et al. (2019a) showed that, on average, players who played more than two rounds of the game (24 picture pairs) chose the right water level class more often than players who played fewer rounds. Players who played more than four rounds (48 picture pairs) were even more accurate.

5.2 Advantages and disadvantages of using an online citizen science game for training

The primary goal of the CrowdWater game is quality control of the crowdsourced data by the citizen scientists themselves. This method has proven successful in improving the quality of the water level class data (Strobl et al., 2019a). The idea to use the game also for training developed over time (see Sect. 2.3). By using an online game for this dual purpose (quality control and training), less effort from project administrators is needed compared to developing a separate online training module and quality control mechanism. Newman et al. (2010) developed multimedia tutorials for a species identification citizen science project and pointed out that they "found the development of multimedia tutorials difficult and time-consuming." (Newman et al., 2010, p. 284).

The CrowdWater game goes beyond the separation of data quality control into "training before the task" and "checking after the task" (Freitag et al., 2016). Instead, training and checking are combined in a continuous loop, where new citizen scientists train and more experienced citizen scientists check the data with the same task. This, in turn, converts new

citizen scientists into more experienced ones after only a few rounds of playing the game. This is similar to iSpot, where citizen scientists upload a picture of a species and identify the species, which is then checked online by other contributors (Silvertown et al., 2015). This leads to the new citizen scientists learning more about species, which will, in turn, make them better at helping other citizen scientists in the future. The approach by Bonter and Cooper (2012) for the FeederWatch project also combined data quality control with training by sending an automatic message to the contributor when a rare and possibly unlikely entry was submitted. They state that these "messages may function as training tools by encouraging participants to become more knowledgeable" (Bonter and Cooper, 2012, p. 306). However, the CrowdWater game is different from these projects in that it does not provide factual knowledge (e.g. on streams or hydrology).

The inclusion of new (and therefore inexperienced) citizen scientists in the quality control process did not negatively influence the quality of the data, mainly due to the averaging of votes of several players (Strobl et al., 2019a). Of course, this is only the case if there are enough experienced players, as well. In the project iSpot, the issue of including beginners in the validation process was solved through reputation scores, which need to be earned through correct species suggestions (Silvertown et al., 2015). This could also be a next step for the CrowdWater game, where an accuracy score can be calculated for each citizen scientist, which can then be used to weight the water level class votes in the game (Strobl et al., 2019a). However, the fact that four rounds of playing the game seem sufficient for training suggests that this is not necessary because new game players quickly turn into experienced ones.

If a citizen science project wants to develop a training task (as opposed to a quality control methodology that also works as a training task), slightly different approaches might be better. In our case, providing the essential criteria for placing a staff gauge in a suitable way (e.g. in between the picture pairs) might have been helpful. Similarly, feedback about the correct water level class could be given directly after each picture pair, rather than after each round of the game (as it is currently implemented). However, this would likely disturb the frequent players. Consequently, our primary goal is data quality control, and most game players are already aware of these criteria and do not want to be disrupted after every picture when they play the game. Therefore, we decided not to add this information to the game. However, additional material, such as tutorial videos, a manual including examples of good and bad staff gauge placements, and introductory app slides are available on the project website. However, our personal experience is that many citizen scientists do not look at this material before using the app and are often not aware of it. A potential benefit of the game, compared to the other material, is that citizen scientists are less likely to see it as "homework" but more as an entertaining activity and are, therefore, likely to spend more time with the game than they would do with other information materials. Encouragingly, participants of this study enjoyed playing the game, meaning that they would participate for the fun aspect instead of seeing it as a "learning task". Consequently, the game can be recommended to any potential citizen scientist, without first having to assess their skills, i.e. their need for training. Additionally, we can recommend that new users play the game instead of discouraging them by explaining that their observations are incorrect.

Citizen science project tasks and therefore also training tasks should always be designed "with the skill of the citizens in mind" (Aceves-Bueno et al., 2017, p. 287). In this study, a similar number of participants rated the game as easy, neutral or difficult. This gives the impression that the difficulty of the game is at a reasonably good level, as it is meant to be engaging and exciting but at the same time not too challenging as to hinder participation. It should be noted that the participants in this study looked at 50 picture pairs in a row in order to simulate several rounds of the regular CrowdWater game, which only shows 12 picture pairs per day. The CrowdWater game itself is, therefore, likely even more accessible because it is less time consuming (and tiring) for citizen scientists.

In the future, it might be feasible to require participants to play the game before starting a new water level class measurement location, thus placing a virtual staff gauge in the CrowdWater app. This would be easily verifiable, as the app and game accounts are the same. In contrast, it is difficult to assess if citizen scientists have read through the introductory slides on the app or the training material that are offered online. Having a compulsory task before all features of the CrowdWater app are available might heighten the barrier to entry, which most citizen science projects that require many participants try to avoid. However, it could also be argued that participants who chose to complete a training session might be more committed towards a project and might, therefore, become more reliable long-term citizen scientists.

5.3 Does participants' self-assessment of confidence predict performance?

In general, participants were more confident in their performance and thought that the task was easier after the training. Self-assessment, however, seems to be an unreliable proxy for actual performance and should, therefore, be interpreted carefully. Participants with a low score for placing or rating the virtual staff gauges might not have realized what the essential criteria were (hence the low score) and therefore also did not realize that their staff gauge placement or rating of the reference pictures was not ideal. Self-assessment might improve after a while, once participants are more aware of which criteria to look for. Such a realization was seen by a CrowdWater app user, who commented that new observations were relatively difficult because the virtual staff gauge in the reference picture that he had created several months earlier was not placed ideally. This indicates that the se-

quence of activities in the CrowdWater project is not ideal, as volunteers have to start with the most difficult part, without having been confronted with different staff gauge placement options. It also suggests that after a while, citizen scientists learn what criteria to look out for and that training may be useful.

The predictability of performance based on self-assessment seems to vary for other studies. McDonough et al. (2017) found that the self-assessed species identification skills did not correspond to the skills of the citizen scientists. Starr et al. (2014) identified a group of citizen scientists who seemed too confident in their abilities but overall believed that the self-assessment was accurate for the majority of their citizen scientists. Crall et al. (2011) found that citizen scientists' skills increased with their self-assessed comfort level. Further research would be required to determine when self-assessment is a reliable prediction of performance. In the meantime, self-assessments should not be fully relied on nor used as a proxy for data quality.

5.4 Limitations of the study

The study was standardized by providing a number of pictures of the same stream to the participants to make the rating of their staff gauge placement comparable and independent of their ability to find a suitable stream. We included a wide range of stream pictures, including some unsuitable angles. The staff gauge placement was assessed for only one river, but it is encouraging to see that there was no difference in the performance of placing the staff gauge after the training online and outdoors, indicating that the online interface and the app were equally intuitive and that participants could also find suitable stream sections on their own. The training, therefore, seems to be teaching the necessary skills to the participants.

Participants could choose from the same 18 stream pictures before and after the training, which could potentially lead to a confirmation bias; i.e. participants might be more likely to choose the same picture after the training as they did before the training. We believe that this effect was negligible, as only two participants with a poor choice of the stream picture before the training still had a poor score after the training as well. All other participants either changed the picture or had already chosen a suitable picture before the training.

By singling out participants with poor performance before the training, the natural variation in performance might lead to improved performance after the training due to a regression towards the mean. However, the improvements were statistically significant when analysed for all participants as well. Further research should investigate how many rounds of the game would be optimal for training the average citizen scientist and if more rounds would lead to better performance for the participants who still received low scores after the training, i.e. if the optimal number of rounds should be adapted depending on the citizen scientist.

A disproportionately large number of study participants in the study had a university degree (85 %) due to the bias in the social network of the authors, recruitment at the university, a tendency of people being more interested in university studies if they have been to university themselves and the study being conducted in English. Many other citizen science projects also report higher participation of university-educated citizen scientists (Brossard et al., 2005; Crall et al., 2011; Overdevest et al., 2004), indicating that the participants of this study might not be that different from the actual citizen scientists in the CrowdWater project.

6 Conclusions

We investigated the value of an online game as a training tool for the CrowdWater project. This game was initially designed for data quality control but turned out to be valuable for improving the participants' ability to set up new observation locations as well. Our results are encouraging beyond the CrowdWater project, and we argue that the overall conclusions that (1) games can provide a suitable approach for training and (2) training and data quality control can be combined also apply to other citizen science projects. Based on our study, the following conclusions about games for training in citizen science projects can be made:

- Citizen science projects should, if possible, be kept intuitive and easy, as this lowers the barrier to entry and might prevent misunderstandings. For the placement of the virtual staff gauge in the CrowdWater project, 70 % of the participants of this study already did well before receiving any training. This compares well with the approximately 10 % error rate for data submitted through the app (Seibert et al., 2019a).
- Games facilitate the training of new citizen scientists and people who have already participated for a while. A big advantage is that this approach is scalable. Large projects with a lot of beginners are also likely to have a lot of advanced citizen scientists, and therefore the number of people who can be trained is not limited by the available time of the people managing the project.
- Training through a game might not necessarily be perceived as training by the citizen scientists (in our case, the primary goal is data quality control). Potentially this helps to make the training feel less like homework before starting to collect data. Nearly two thirds of the participants of this training study said that the game was fun; this compares well with a survey among early game players of whom 86 % said that they enjoyed playing the game (Strobl et al., 2019a).
- While materials such as manuals and tutorials can be useful, gamified approaches provide an enjoyable al-

ternative training mechanism for citizen scientists. Citizen scientists might respond differently to various training techniques. In our case, we noticed that few citizen scientists read the manual or watched the instruction videos but also that some individuals might have responded better to a more explicit and less playful training method. We, therefore, recommend offering different training options.

Data availability. All data files are available from the Zenodo data repository (https://doi.org/10.5281/zenodo.3538008; Strobl et al., 2019b).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/gc-3-109-2020-supplement.

Author contributions. BS conceptualized the project, curated the data, performed the formal analysis, led the investigation, planned the methodology, administered the project, procured the resources, visualized the data and wrote the original draft of this paper. SE conceptualized the project, planned the methodology, and reviewed and edited the paper. HJIvM conceptualized the project, acquired funding, planned the methodology, supervised the project, and reviewed and edited the paper. JS conceptualized the project, acquired funding, planned the methodology, supervised the project, and reviewed and edited the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank all 52 study participants for their time and interest in this study. We also thank all CrowdWater app and CrowdWater game contributors for providing the database that was used in this study and for submitting valuable hydrological data. We hope that the participants will continue to participate in the CrowdWater project by using the app or playing the game and invite all readers to use the app and/or play the game as well.

Financial support. This research has been supported by the Swiss National Science Foundation (grant no. 200021_163008).

Review statement. This paper was edited by Chris King and reviewed by two anonymous referees.

References

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., and Anderson, S. E.: The Accuracy of Citizen Science Data: A Quantitative Review, Bull. Ecol. Soc. Am., 98, 278–290, https://doi.org/10.1002/bes2.1336, 2017.

Barras, H., Hering, A., Martynov, A., Noti, P.-A., Germann, U., and Martius, O.: Experiences with >50'000 crowd-sourced hail reports in Switzerland, Am. Meteorol. Soc., 100, 1429–1440, https://doi.org/10.1175/bams-d-18-0090.1, 2019.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J. and Wilderman, C. C.: Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education, Washington, DC, available at: https://files.eric.ed.gov/fulltext/ED519688.pdf (last access: 30 April 2020), 2009.

Bonney, R., Phillips, T. B., Ballard, H. L., and Enck, J. W.: Can citizen science enhance public understanding of science?, Public Underst. Sci., 25, 2–16, https://doi.org/10.1177/0963662515607406, 2016.

Bonter, D. N. and Cooper, C. B.: Data validation in citizen science: A case study from Project FeederWatch, Front. Ecol. Environ., 10, 305–307, https://doi.org/10.1890/110273, 2012.

Breuer, L., Hiery, N., Kraft, P., Bach, M., Aubert, A. H., and Frede, H.-G.: HydroCrowd: a citizen science snapshot to assess the spatial control of nitrogen solutes in surface waters, Sci. Rep., 5, 16503, https://doi.org/10.1038/srep16503, 2015.

Brossard, D., Lewenstein, B., and Bonney, R.: Scientific knowledge and attitude change: The impact of a citizen science project, Int. J. Sci. Educ., 27, 1099–1121, https://doi.org/10.1080/09500690500069483, 2005.

Canfield, D. E., Bachmann, R. W., Stephens, D. B., Hoyer, M. V., Bacon, L., Williams, S., and Scott, M.: Monitoring by citizen scientists demonstrates water clarity of Maine (USA) lakes is stable, not declining, due to cultural eutrophication, Inl. Waters, 6, 11–27, https://doi.org/10.5268/IW-6.1.864, 2016.

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., and Waller, D. M.: Assessing citizen science data quality: An invasive species case study, Conserv. Lett., 4, 433–442, https://doi.org/10.1111/j.1755-263X.2011.00196.x, 2011.

Crall, A. W., Jordan, R., Holfelder, K., Newman, G. J., Graham, J., and Waller, D. M.: The impacts of an invasive species citizen science training program on participant attitudes, behavior, and science literacy, Public Underst. Sci., 22, 745–764, https://doi.org/10.1177/0963662511434894, 2013.

Cronje, R., Rohlinger, S., Crall, A., and Newman, G.: Does Participation in Citizen Science Improve Scientific Literacy? A Study to Compare Assessment Methods, Appl. Environ. Educ. Commun., 10, 135–145, https://doi.org/10.1080/1533015X.2011.603611, 2011.

Dem, E. S., Rodríguez-Labajos, B., Wiemers, M., Ott, J., Hirneisen, N., Bustamante, J. V., Bustamante, M., and Settele, J.: Understanding the relationship between volunteers' motivations and learning outcomes of Citizen Science in rice ecosystems in the Northern Philippines, Paddy Water Environ., 16, 725–735, https://doi.org/10.1007/s10333-018-0664-9, 2018.

Etter, S., Strobl, B., Seibert, J., and van Meerveld, I.: Value of uncertain streamflow observations for hydrological modelling, Hydrol. Earth Syst. Sci., 22, 5243–5257, https://doi.org/10.5194/hess-22-5243-2018, 2018.

Evans, C., Abrams, E., Reitsma, R., Roux, K., Salmonsen, L., and Marra, P. P.: The Neighborhood Nestwatch Program: Participant Outcomes of a Citizen-Science Ecological Research Project, Conserv. Biol., 19, 589–594, https://doi.org/10.1111/j.1523-1739.2005.00s01.x, 2005.

Freitag, A., Meyer, R., and Whiteman, L.: Strategies Employed by Citizen Science Programs to Increase the Credibility of Their Data, Citiz. Sci. Theory Pract., 1, 1–11, https://doi.org/10.5334/cstp.6, 2016.

Gaddis, M.: Training Citizen Scientists for Data Reliability: a Multiple Case Study to Identify Themes in Current Training Initiatives, University of the Rockies, 2018.

Goodchild, M. F.: Citizens as sensors: The world of volunteered geography, GeoJ., 69, 211–221, https://doi.org/10.1007/s10708-007-9111-y, 2007.

Jennett, C., Kloetzer, L., Schneider, D., Iacovides, I., Cox, A. L., Gold, M., Fuchs, B., Eveleigh, A., Mathieu, K., Ajani, Z., and Talsi, Y.: Motivations, learning and creativity in online citizen science, J. Sci. Commun., 15, 1–23, 2016.

Jordan, R. C., Gray, S. A., Howe, D. V., Brooks, W. R., and Ehrenfeld, J. G.: Knowledge Gain and Behavioral Change in Citizen-Science Programs, Conserv. Biol., 25, 1148–1154, https://doi.org/10.1111/j.1523-1739.2011.01745.x, 2011.

Kampf, S., Strobl, B., Hammond, J., Annenberg, A., Etter, S., Martin, C., Puntenney-Desmond, K., Seibert, J., and van Meerveld, I.: Testing the waters: Mobile apps for crowdsourced streamflow data, Eos (Washington, DC), 30–34, https://doi.org/10.1029/2018EO096355, 2018.

Kremen, C., Ullman, K. S., and Thorp, R. W.: Evaluating the Quality of Citizen-Scientist Data on Pollinator Communities, Conserv. Biol., 25, 607–617, https://doi.org/10.1111/j.1523-1739.2011.01657.x, 2011.

Krennert, T., Kaltenberger, R., Pistotnik, G., Holzer, A. M., Zeiler, F., and Stampfl, M.: Trusted Spotter Network Austria – a new standard to utilize crowdsourced weather and impact observations, Adv. Sci. Res., 15, 77–80, https://doi.org/10.5194/asr-15-77-2018, 2018.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., and Vandenberg, J.: Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey, Mon. Not. R. Astron. Soc., 389, 1179–1189, https://doi.org/10.1111/j.1365-2966.2008.13689.x, 2008.

Little, K. E., Hayashi, M., and Liang, S.: Community-Based Groundwater Monitoring Network Using a Citizen-Science Approach, Groundwater, 54, 317–324, https://doi.org/10.1111/gwat.12336, 2016.

Lowry, C. S., Fienen, M. N., Hall, D. M., and Stepenuck, K. F.: Growing Pains of Crowdsourced Stream Stage Monitoring Using Mobile Phones: The Development of CrowdHydrology, Front. Earth Sci., 7, 1–10, https://doi.org/10.3389/feart.2019.00128, 2019.

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., Padmanabhan, S., Nielsen, K., and Ozcan, A.: Distributed medical image analysis and diagnosis through crowdsourced games: A malaria case study, PLoS One, 7, 1–8, https://doi.org/10.1371/journal.pone.0037245, 2012.

McDonough MacKenzie, C., Murray, G., Primack, R., and Weihrauch, D.: Lessons from citizen science: Assessing volunteer-collected plant phenology data with Mountain Watch, Biol. Conserv., 208, 121–126, https://doi.org/10.1016/j.biocon.2016.07.027, 2017.

Newman, G., Crall, A., Laituri, M., Graham, J., Stohlgren, T., Moore, J. C., Kodrich, K., and Holfelder, K. A.: Teaching citizen science skills online: Implications for invasive species training programs, Appl. Environ. Educ. Commun., 9, 276–286, https://doi.org/10.1080/1533015X.2010.530896, 2010.

Overdevest, C., Orr, C. H., and Stepenuck, K.: Volunteer Stream Monitoring and Local Participation in Natural Resource Issues, Hum. Ecol. Rev., 11, 177–185, 2004.

Phillips, T. B., Ballard, H. L., Lewenstein, B. V., and Bonney, R.: Engagement in science through citizen science: Moving beyond data collection, Sci. Educ., 103, 665–690, https://doi.org/10.1002/sce.21501, 2019.

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., Raven, P. H., Roberts, C. M., and Sexton, J. O.: The biodiversity of species and their rates of extinction, distribution, and protection, Science, 344, 6187, https://doi.org/10.1126/science.1246752, 2014.

Reges, H. W., Doesken, N., Turner, J., Newman, N., Bergantino, A., and Schwalbe, Z.: CoCoRaHS: The Evolution and Accomplishments of a Volunteer Rain Gauge Network, Am. Meteorol. Soc., 97, 1831–1846, https://doi.org/10.1175/BAMS-D-14-00213.1, 2016.

Rinderer, M., Komakech, H. C., Müller, D., Wiesenberg, G. L. B., and Seibert, J.: Qualitative soil moisture assessment in semi-arid Africa – the role of experience and training on inter-rater reliability, Hydrol. Earth Syst. Sci., 19, 3505–3516, https://doi.org/10.5194/hess-19-3505-2015, 2015.

Rufino, M. C., Weeser, B., Stenfert-Kroese, J., Njue, N., Gräf, J., Jacobs, S., Kemboi, Z., Ran, A. M., Cerutti, P. O., Martius, C., and Breuer, L.: Citizen scientists monitor water quantity and quality in Kenya, CIFOR infobriefs, 230, 1–4, https://doi.org/10.17528/cifor/007013, 2018.

Seibert, J., Strobl, B., Etter, S., Hummer, P., and van Meerveld, H. J.: Virtual Staff Gauges for Crowd-Based Stream Level Observations, Front. Earth Sci., 7, 70, https://doi.org/10.3389/feart.2019.00070, 2019a.

Seibert, J., van Meerveld, H. J., Etter, S., Strobl, B., Assendelft, R., and Hummer, P.: Wasserdaten sammeln mit dem Smartphone – Wie können Menschen messen, was hydrologische Modelle brauchen?, Hydrol. Wasserbewirts., 63, 74–84, https://doi.org/10.5675/HyWa_2019.2_1, 2019b.

Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J., and McConway, K.: Crowdsourcing the identification of organisms: A case-study of iSpot, Zookeys, 480, 125–146, https://doi.org/10.3897/zookeys.480.8803, 2015.

Starr, J., Schweik, C. M., Bush, N., Fletcher, L., Finn, J., Fish, J., and Bargeron, C. T.: Lights, Camera-Citizen Science: Assessing the Effectiveness of Smartphone-Based Video Training in Invasive Plant Identification, PLoS One, 9, 1–7, https://doi.org/10.1371/journal.pone.0111433, 2014.

Strobl, B., Etter, S., van Meerveld, I., and Seibert, J.: The CrowdWater game: A playful way to improve the accuracy of crowdsourced water level class data, edited by: Mirjalili, S., PLoS One, 14, e0222579, https://doi.org/10.1371/journal.pone.0222579, 2019a.

Strobl, B., Etter, S., van Meerveld, I. H. J., and Seibert, J.: CrowdWater game training study, Zenodo, https://doi.org/10.5281/zenodo.3538008, 2019b.

van Meerveld, H. J., Vis, M. J. P., and Seibert, J.: Information content of stream level class data for hydrological

model calibration, Hydrol. Earth Syst. Sci., 21, 4895–4905, https://doi.org/10.5194/hess-21-4895-2017, 2017.

Weeser, B., Stenfert Kroese, J., Jacobs, S. R., Njue, N., Kemboi, Z., Ran, A., Rufino, M. C., and Breuer, L.: Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring, Sci. Total Environ., 631/632, 1590–1599, https://doi.org/10.1016/j.scitotenv.2018.03.130, 2018.